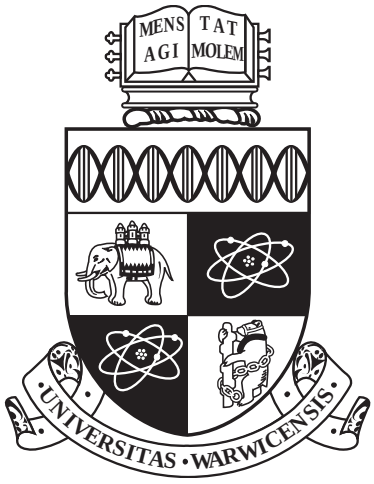


Applications of Entropy to Extremal Problems

by

Matthew Fitch

Thesis

Submitted to the University of Warwick

for the degree of

Doctor of Philosophy

Department of Mathematics

September 2018

THE UNIVERSITY OF
WARWICK

Contents

Acknowledgments

Thanks to my supervisor, Oleg Pikhurko, for help and guidance.

Thanks to the European Research Council, who funded me with Grant No. 306493.

Declarations

I have put the following papers on Arxiv, which contain the same material as can be found here:

Rational exponents for hypergraph Turán problems, Matthew Fitch, arXiv:1607.05788

Kruskal-Katona type Problem, Matthew Fitch, arXiv:1805.00340

Implicit representation conjecture for semi-algebraic graphs, Matthew Fitch, arXiv:1803.01882

Abstract

The Sidorenko conjecture gives a lower bound on the number of homomorphisms from a bipartite graph to another graph. Szegedy [28] used entropy methods to prove the conjecture in some cases. We will refine these methods to also give lower bounds for the number of injective homomorphisms from a bipartite graph to another bipartite graph, and a lower bound for the number of homomorphisms from a k-partite hypergraph to another k-partite hypergraph, as well as a few other similar problems.

Next is a generalisation of the Kruskal Katona Theorem [19, 17]. We are given integers $k < r$ and families of sets $\mathcal{A} \subset \mathbb{N}^{(r)}$ and $\mathcal{B} \subset \mathbb{N}^{(r-1)}$ such that for every $A \in \mathcal{A}$, at least k distinct subset of A of size $r-1$ are in $\mathcal{B}$. We then ask the question of what is the minimum size of $\mathcal{A}$ as a function of the size of $\mathcal{B}$? In the case where $0 \leq k \leq 3$, we will be able to find an exact solution. Then for $k \geq 4$ we will make a lot of progress towards finding a solution.

The next chapter is to do with Turán-type problems. Given a family of k-hypergraphs $\mathcal{F}$, $ex(n, \mathcal{F})$ is the maximum number of edges an $\mathcal{F}$-free n-vertex k-hypergraph can have. We prove that for a rational r, there exists some finite family $\mathcal{F}$ of k-hypergraphs for which $ex(n, \mathcal{F}) = \Theta(n^{k-r})$ if and only if $0 \leq r \leq k-1$ or $r = k$.

The final chapter will deal with the implicit representation conjecture, in the special case of semi-algebraic graphs. Given a graph in such a family, we want to assign a name to each vertex in such a way that we can recover each edge based only on the names of the two incident vertices. We will first prove that one 'obvious' way of storing the information doesn't work. Then we will come up with a way of storing the information that requires $O(n^{1-\epsilon})$ bits per vertex, where ϵ is some small constant depending only on the family.

Chapter 1

Introduction

1.1 Injectivity-like property added to the Sidorenko conjecture

Chapter 2 is about a result related to the Sidorenko conjecture. The Sidorenko conjecture was asked by Sidorenko [24], and also by Erdos and Simonovits [26]. It states that if G is a bipartite graph with $e(G)$ edges and X is a graph with n vertices and average degree d, then the number of homomorphisms from G to X is at least $nd^{e(G)}$. The conjecture has not been solved completely, but has for a large number of cases, including paths, by Blackley and Roy [5], cycles, trees and complete bipartite graphs by Sidorenko [23], and a few more exotic cases. In particular, Szegedy proved it for a very general class of graphs [28] and Szegedy's proof will be the one most useful to us, in particular when it is applied to trees.

The Sidorenko conjecture is interesting because many problems in combinatorics require bounding subgraphs, and the Sidorenko conjecture, in the cases where it has been proved, is a powerful tool for doing so. However, one complication when using it is that the image of a homomorphism from G to X is not an actual copy of G: it can self-intersect. The most extreme example of this is that since G is bipartite, there exists a homomorphism that sends each side of the bipartition of G to a vertex, so therefore a single edge is an image of this homomorphism. Usually in applications, we want to exclude such things and require that the copies of H don't self-intersect, or in other words, we want the homomorphisms to be injective.

What we are doing here is we are first going to simply present Szegedy's proof in the

case where G is a tree. This is because we will use elements and intermediate results from it to do the proof of the more general result. That is to say, we will provide a lower bound for the number of injective homomorphisms from G to X. In fact, we will be slightly more general, and count the number of homomorphisms with an 'injective-like' property $\mathcal{P}$. The result we get is that as $n, d \to \infty$, the number of homomorphisms with property $\mathcal{P}$ is at least $cnd^{e(G)}(1 - o(1))$, where $c \leq 1$ is a constant that we give explicitly. The main reason for doing this generalisation is because we want to use it in two other problems, which include some injective-like properties for copies of trees. These two problems were actually the principal motivation for developing this modification to Szegedy's proof. We will present these two applications in sections 3 and 4.

1.2 A Kruskal-Katona type problem

The third chapter is about a variant on the Kruskal-Katona Theorem. In the Kruskal Katona Theorem, we are given families of sets $\mathcal{A} \subset [n]^{(r)}$ and $\mathcal{B} \subset [n]^{(r-1)}$ such that for every $A \in \mathcal{A}$, there exist distinct $B_1, B_2, ..B_r \in \mathcal{B}$ with $B_i \subset A$. Given $b = |\mathcal{B}|$, we want to maximise $a = |\mathcal{A}|$. The Kruskal-Katona Theorem states that an optimal solution is when both $\mathcal{A}$ and $\mathcal{B}$ are initial segments of the colexicographic ordering on sets of size r and $r-1$ respectively.

We want to consider the modified problem, posed by Bollobás and Eccles [6] where instead of requiring all r subsets of A of size $r-1$, we only require k of them for some $k < r$: given families of sets $\mathcal{A} \subset [n]^{(r)}$ and $\mathcal{B} \subset [n]^{(r-1)}$ such that for every $A \in \mathcal{A}$, there exist distinct $B_1, B_2, ..B_k \in \mathcal{B}$ with $B_i \subset A$. Given $b = |\mathcal{B}|$, we want to maximise $a = |\mathcal{A}|$.

The conjectured solution is that there exists some set S of size $r-k$ and then $\mathcal{A}$ consists of sets in an initial segment of colex on sets of size k that are then unioned with S. $\mathcal{B}$ is similarly comprised of sets in an initial segment of colex on sets of size $k-1$ that are then unioned with S.

We observe that this conjecture is true in the cases $k = 0$, $k = 1$, $k = 2$. We also prove the case $k = 3$ as long as $b \geq 406$. For the case $k \geq 4$, we prove the result for infinitely many values of b but not all. We also prove that this conjecture is within an additive constant of the real result.

The first few subsections are dealing with the trivial cases of $k = 0$ and $k = 1$ fol-

lowed by the less trivial but still pretty straightforward case $k = 2$.
The next subsection deals with the case $k = 3$. A very rough idea about how the proof goes is as follows: We first relate the values of a and b by counting the number of paths of length 2 (a path of length 2 means 2 elements of $\mathcal{A}$ and 3 elements of $\mathcal{B}$ in a sequence where two successive sets differ only by the addition or subtraction of a single element). This proves the conjecture in the cases where $|\mathcal{B}| = \binom{c}{2}$ for some integer c. For other values of $|\mathcal{B}|$, we use stability to say things about a potential counter-example and eventually prove that any counter-example must 'look' a lot like the solution to the classical version of Kruskal Katona. We can then apply the classical version of Kruskal Katona to get even better bounds on a, and these are so tight that we end up having only 3 potential classes of counter examples. We finish up by looking at each of these 3 in turn and realise none of them are possible. Therefore the conjecture holds for all $b \geq 406$.

The final subsection deals with the case $k \geq 4$ and this is where we use the modified version of Sidorenko from section 2. Unfortunately, this section contains a gap in the proof which I only noticed late and which hasn't been fixed. We make a conjecture which if true, fixes the issue. Assuming this conjecture is true, the rest of the proof is very similar to the case $k = 3$ but a few additional complications arise. A rough outline is as follows: we start by bounding the number of paths of length $k - 1$ using our Sidorenko result. This allows us to give some good bounds on a and b. This doesn't immediately give us a sharp bound like in the case $k = 3$, but it is still close enough to allow us to use stability and it gives information about what a potential counter-example would look like. It also 'looks' a lot like a solution to the classical version of Kruskal Katona, which is enough to let us apply the classical Kruskal Katona Theorem to get even better bounds on a. These end up being so tight that for any k, that we can prove the conjecture for infinitely many b. It also tells us that for every k, there exists some constant $\tau(k)$ and then the real optimal value for a has to be within $\tau(k)$ of the conjectured optimal value for a.

1.3 Rational Exponents for hypergraph Turán problems

The fourth chapter will be about the following result:

Given a family of k-hypergraphs $\mathcal{F}$, $ex(n, \mathcal{F})$ is the maximum number of edges a k-hypergraph can have, knowing that said hypergraph has n vertices but contains no copy of any hypergraph from $\mathcal{F}$ as a subgraph. We prove that for a rational r, there

exists some finite family $\mathcal{F}$ of k-hypergraphs for which $ex(n, \mathcal{F}) = \Theta(n^{k-r})$ if and only if $0 \leq r \leq k-1$ or $r = k$,.

Finding $ex(n, \{F\})$ for a fixed graph or hypergraph F is known as the Turán problem. For ordinary ($k = 2$) non-bipartite graphs, we have a reasonable understanding: Turán gave an exact solution when F is a complete graph [29], while Erdős and Stone gave an asymptotic solution for any non-bipartite graph [12]: when H is a graph with n vertices and chromatic number $\chi(H)$,

$$ex(n, H) = \left(1 - \frac{1}{\chi(H) - 1} + o(1)\right) \binom{n}{2}.$$

However, for bipartite graphs and more general hypergraphs ($k \geq 3$), very little is known, not even asymptotically [25, 18]. For a lot of families of k-hypergraphs, $ex(n, \mathcal{F})$ is of order $\Omega(n^k)$. However, there are some for which $ex(n, \mathcal{F})$ is of order $o(n^k)$. We call this case a *degenerate Turán problem*. Erdős [9] found some bounds for the extremal number for the complete k-partite k-hypergraph with equal partitions of size l: there exists a constant C such that for all n sufficiently large,

$$n^{k - C/l^{k-1}} < ex(n, K^{(k)}(l, l, l..., l)) \leq n^{k - 1/l^{k-1}}.$$

This implies that whenever $\mathcal{F}$ contains a k-partite k-hypergraph, this is a degenerate Turán problem.

In 1979, Erdős [10] conjectured that for every rational r between 1 and 2, there exists a finite family of bipartite graphs $\mathcal{F}$ with $ex(n, \mathcal{F}) = \Theta(n^r)$. This conjecture was later proved in 2015 by Bukh and Conlon [8].

In 1986, Frankl [13] proved a related result for hypergraphs: for every rational $r \geq 1$, there exists some $k \in \mathbb{N}$ and some finite family $\mathcal{F}$ of k-hypergraphs such that $ex(n, \mathcal{F})$ is of order n^r. (Side-note: the $\mathcal{F}$ that Frankl used also had the property that every $F \in \mathcal{F}$ had exactly 2 edges.) .

In 2016, Ma, Yuan and Zhang [21] discovered an infinite family of k-hypergraphs for which they could solve the Turán problem asymptotically. They proved that $K^{(k)}_{s_1, s_2, ..., s_k}$, the complete k-partite k-hypergraph with partition sizes $s_1, s_2, ..., s_k$ has $ex(n, K^{(k)}_{s_1, s_2, ..., s_k}) =$

$\Theta(n^{k-\frac{1}{s_1 s_2 s_3 \cdots s_{k-1}}})$ whenever s_k is sufficiently large..

The way this section is organised is by first constructing the set of hypergraphs that will work for $0 \leq r < 1$. The construction is similar to that from [8]. The second subsection is dedicated to proving that our construction satisfies the lower bound. We use similar techniques to those used in [8] when they proved their lower bound. The third subsection deals with proving the upper bound, and this is where our Sidorenko result comes in. This ends up proving the upper bound in the case $0 \leq r < 1$. The last subsection deals with generalising the result from $0 \leq r < 1$ to $0 \leq r < k-1$. We also show that $r = k-1$ and $r = k$ are both possible, but $k-1 < r < k$ is impossible.

1.4 Implicit representation conjecture for semi-algebraic graphs

The fourth and final section is of a slightly different flavour.

A semi-algebraic family of graphs consists of a Euclidean space $\mathcal{S}$ and a set of polynomial equalities and inequalities on $\mathcal{S} \times \mathcal{S}$, the set of graphs in the family are exactly those graphs whose vertices are points in $\mathcal{S}$ and whose edges are exactly those pairs of points that satisfy all the polynomial equalities and inequalities.

For each vertex of a graph H with n vertices, we associate $m(n)$ bits of information. That means there is a sequence of functions F_n such that for every integer n, F_n is a function from the set of graphs with n vertices to $[2^{m(n)}]^n$ and there there is a function G_n from $[2^{m(n)}] \times [2^{m(n)}]$ to $\{0,1\}$ such that for every pair of vertices x and y of H, $G_n(F_n(H)_x, F_n(y)_y) = 1$ if and only if xy is an edge of H. The problem we want to solve is to minimise m. This was posed in [16, 27].

After a short subsection detailing a few simple results about semi-algebraic graphs, we prove that a 'natural' hypothesis for F and G doesn't actually work. This hypothesis is to approximate the coordinates of all the vertices using algebraic numbers and then just store these algebraic numbers. This builds upon the work by McDiarmid and Müller [22], who proved that storing approximations to the coordinates as integers didn't work because there exists a sequence of graphs in a semi-algebraic family for which the maximum crossratio $|a-b|/|c-d|$ is too large for the vertices to be stored using integers.

Kang and Müller [15] improved upon this and showed that this crossratio was also too large for the vertices to be stored using rational numbers, and that therefore storing rational approximations of the coordinates of the vertices doesn't work. Our proof shows that this crossratio is too large for the vertices to be stored using algebraic numbers.

The last subsection is about a completely different method, and we show that it does improve upon the previous best known bound. It is loosely inspired by some of the methods used in [2] to prove properties about algebraic graphs. Our method improves the upper bound from $n/2 + \log_2(n)$ (the trivial bound) to $cn^{1-\epsilon}$ for some small constant $\epsilon > 0$ and some constant c. This is still far from the lower bound and conjectured solution of $O(\log_2(n))$ but it is still a small improvement.

1.4.1 Definitions and notation

Here we collect some notation we will often use.

Definition 1. *Given an integer $k \geq 2$, a* k-hypergraph *G is a set of points (called the* vertices*), together with a collection of k-subsets of the vertices (called the* edges*). For such a k-hypergraph, let $V(G)$ mean its vertex set, and $E(G)$ its edge set. Also define $|G| = v(G) = |V(G)|$ to be its number of vertices and $e(G) = |E(G)|$ to be the number of edges.*

Given two hypergraphs G and G', their union $G \cup G'$ is defined to be the hypergraph that has $V(G) \cup V(G')$ as its vertex set, and $E(G) \cup E(G')$ as its edge set.

We say that $G' \subset G$ if $V(G') \subset V(G)$ and $E(G') \subset E(G)$.

Definition 2. *Given k-hypergraphs G and X, a* graph homomorphism *(often shortened to* homomorphism*) from G to X means a function f that assigns to each vertex of G some vertex in X and that also preserves edges, i.e. for every edge $\{x_1, x_2..., x_k\}$ in G, $\{f(x_1), f(x_2), ..., f(x_k)\}$ is also an edge of X.*

The set of homomorphisms from G to X is denoted by $\mathrm{Hom}(G, X)$.

Definition 3. *Given a homomorphism $H \in \mathrm{Hom}(G, X)$ and a subgraph $G' \subset G$, the restriction of H to G' is the homomorphism $H' \in \mathrm{Hom}(G', X)$ defined by $H'(x) = H(x)$*

for every vertex $x \in G'$.
In this case, we also call H an extension *of H' from G' to G.*

Definition 4. *A property $\mathcal{P}$ of homomorphisms is a function from $Hom(G, X)$ to $\{true, false\}$ for some G and X.*
If $\mathcal{P}$ is a property of homomorphisms, then we will denote by $\mathrm{Hom}_{\mathcal{P}}(G, X)$ the set of homomorphisms that satisfy property $\mathcal{P}$.

For example, 'injectivity' is a property of homomorphisms from G to X for any choice of G and X. A homomorphism $H \in \mathrm{Hom}(G, X)$ is injective if and only if $H(x) \neq H(y)$ for every choice of vertices $x \neq y$ in G.

Definition 5. *A discrete probability distribution μ on some countable set S is a function from S to the real interval $[0, 1]$ such that $\sum_{s \in S} \mu(s) = 1$.*

Definition 6. *Given a discrete probability distribution μ on a countable set S and a subset A of S, $\mathbb{P}(A) = \sum_{s \in A} \mu(s)$ means the probability that event A will occur.*

Definition 7. *Given a discrete probability distribution μ on a countable set S and a random variable $B : S \to \mathbb{R}$, the expectation of B is $\mathbb{E}(B) = \sum_{s \in S} B(s)\mu(s)$ if it exists.*

Definition 8. *Given a discrete probability distribution μ on a countable set S, the entropy of μ is defined to be $D(\mu) = \sum_{s \in S} -\ln(\mu(s))\mu(s)$ if it exists.*

Lemma 1. *Given a probability distribution μ on a finite set S, we have:*

$$D(\mu) \leq \ln(|S|)$$

Proof of Lemma 1: The function $-x\ln(x)$ is concave on the interval between 0 and 1. This implies that $\frac{D(\mu)}{|S|} = \frac{\sum_{s \in S} -\ln(\mu(s))\mu(s)}{|S|} \leq -\ln\left(\frac{\sum_{s \in S} \mu(s)}{|S|}\right) \frac{\sum_{s \in S} \mu(s)}{|S|} = -\ln\left(\frac{1}{|S|}\right) \frac{1}{|S|} = \frac{\ln(|S|)}{|S|}$, and therefore $D(\mu) \leq \ln(|S|)$.

Chapter 2

Injectivity-like property added to the Sidorenko conjecture

2.1 Introduction

The Sidorenko conjecture states that if G is a bipartite graph with $e(G)$ edges and X is a graph with n vertices and average degree d, then the number of homomorphisms from G to X is at least $nd^{e(G)}$. Here, we will only be looking at the case where $G = T$ is a tree, for which there do exist a number of proofs, most notably the proof by Szegedy [28] which uses entropy.

However, the number of homomorphisms is not that useful in applications because just a single edge on its own is the image of a homomorphism from a bipartite graph. Generally, in applications, we want a bound on the number of certain special types of homomorphisms, for example, we might want a bound on the number of injective homomorphisms. We will denote these 'special' homomorphisms we're interested in by saying that they have some property $\mathcal{P}$. If $\mathcal{P}$ satisfies certain axioms (which we will shorten to saying if $\mathcal{P}$ is injective-like), then we can find a lower bound on the number of our special homomorphisms satisfying $\mathcal{P}$.

To best motivate how we will choose our axioms for being 'injective-like', we will start with an observation about homomorphisms in general. Suppose we have a subtree $T' \subsetneq T$ and a homomorphism $H \in \mathrm{Hom}(T', X)$. Let E be an edge in T that is not in T' but is incident to it, say at some vertex t. Thus, $T' \cup E$ is a slightly larger subtree of T. We want to find a homomorphism $H' \in \mathrm{Hom}(T' \cup E, X)$ that extends H, or in other

words, such that the restriction of H' to T' is exactly H. This means that we just need to find a suitable image for E in X. The only restriction is that $H'(E)$ has to be incident to $H'(t) = H(t)$. Therefore there are exactly $\deg(H(t))$ possible choices for an $H' \in \mathrm{Hom}(T' \cup E, X)$ that extends H.

What happens if, instead of looking at all homomorphisms, we only look at injective ones? What changes in the above observation is that at each step, we lose at most a constant number of homomorphisms. Indeed, when we pick $H'(E)$, we need it to be incident to $H'(t)$, giving $\deg(H(t))$ possible choices, but we also need the other end to not be any vertex in $H(T')$. So we end up with between $\deg(H(t)) - |V(T')|$ and $\deg(H(t))$ possibilities for an $H' \in \mathrm{Hom}(T' \cup E, X)$ that extends H.

The intuitive definition for $\mathcal{P}$ being 'injective-like' will essentially say that if we replace Hom with $\mathrm{Hom}_{\mathcal{P}}$ in the above observation, then we only lose a constant proportion of homomorphisms compared to the regular case. At each step, we will have some real positive constants c_1 and c_2 such that there are $c_1 \deg(H(t)) - c_2$ possibilities for an $H' \in \mathrm{Hom}(T' \cup E, X)$ that extends H. More precisely:

Definition 9. *Given a tree T and a graph X, a property $\mathcal{P}$ of homomorphisms from subtrees of T to X is* weakly injective-like *if there exists some real positive constant $p \geq 0$, and for every subtree $T' \subset T$, there exists a real positive constant $0 < q_{T'} \leq 1$ such that the following all hold:*
• *$\mathcal{P}$ holds for all vertex homomorphisms. In other words, if S is a tree with 1 vertex and no edges, then $\mathrm{Hom}(S, X) = \mathrm{Hom}_{\mathcal{P}}(S, X)$.*
• *If S is a tree with 1 vertex and no edges, then $q_S = 1$.*
• *Given any subtree $T' \subsetneq T$, an edge E incident to T' at t, and a homomorphism $H \in \mathrm{Hom}_{\mathcal{P}}(T', X)$, then the number of ways to extend H to some $H' \in \mathrm{Hom}_{\mathcal{P}}(T' \cup E, X)$ is at least $\deg(H(t))\frac{q_{T' \cup E}}{q_{T'}} - p$.*

Definition 10. *Given a tree T and a graph X, a property $\mathcal{P}$ of homomorphisms from subtrees of T to X is* strongly injective-like *if, in addition to being weakly injective-like, the number of ways to extend H to some $H' \in \mathrm{Hom}_{\mathcal{P}}(T' \cup E, X)$ is at most $\deg(H(t))\frac{q_{T' \cup E}}{q_{T'}}$.*

Example: Let T be an tree, X a graph, and let $\mathcal{P}$ be injectivity. We set $f = v(T) - 1$ and for every subtree $T' \subset T$, we set $q_{T'} = 1$. We check the three bullet points:
• A homomorphism from a single vertex is always injective so the first bullet point holds.

• The second bullet point holds trivially.
• For the third bullet point, suppose we have a homomorphism $H \in \mathrm{Hom}_{\mathcal{P}}(T', X)$, and we want to add a new edge E to T', incident to it at t and then get a new homomorphism $H' \in \mathrm{Hom}_{\mathcal{P}}(T' \cup E, X)$. There are $\deg(H(t))$ possibilities for this new edge to make a homomorphism. However, not all the resultant homomorphisms are injective. For it to be injective, we need the new vertex to NOT be any of the vertices in $H(T')$. There are at most $v(H(T')) - 1 \leq v(T) - 1 = f$ edges that join $H(t)$ to another vertex of $H(T')$ and as long as we avoid those we are fine. So there are at least $\deg(H(t)) - f$ possibilities for $H' \in \mathrm{Hom}_{\mathcal{P}}(T' \cup E, X)$ that extends H so the third bullet point holds. Thus, 'injectivity' is weakly injective-like. But furthermore, the maximum number of possibilities for $H' \in \mathrm{Hom}_{\mathcal{P}}(T' \cup E, X)$ that extends H is $\deg(H(t))$ so injectivity is strongly injective-like.

Theorem 1 (Szegedy [28])**.** *Let X be a graph with n vertices and average degree d and let T be a tree with e edges. Then $|Hom(T, X)| \geq nd^e$.*

Our main result is a modification of this to include a property $\mathcal{P}$:

Theorem 2. *Let T be a tree with e edges and let X be a graph with $n \geq 3$ vertices and average degree d. Suppose $\mathcal{P}$ is a strongly injective-like property with p and $g_{T'}$ as before. Then $|\mathrm{Hom}_{\mathcal{P}}(T, X)| \geq nd^e q_T \cdot \left(1 - e(e+2)(1 + q_T^{-1})\frac{\ln(n)p}{d}\right)$.*

We also conjecture that this also holds when $\mathcal{P}$ is weakly injective-like, though we have been unable to prove it:

Conjecture 1. *Let T be a tree with e edges and let X be a graph with $n \geq 3$ vertices and average degree d. Suppose $\mathcal{P}$ is a weakly injective-like property with p and $g_{T'}$ as before. Then $|\mathrm{Hom}_{\mathcal{P}}(T, X)| \geq nd^e q_T \cdot \left(1 - e(e+2)(1 + q_t^{-1})\frac{\ln(n)p}{d}\right)$.*

How are we going to go about proving Theorem 2? To start, it is useful to look at Szegedy's proof for Theorem 1 because our proof will use a lot of the same elements. Once we have completed the proof of Theorem 1, we can move on to Theorem 2. To give a rough idea of how both proofs work, we are going to proceed by induction on the number of edges of T. The first step is to pick a vertex t_1 in T, and think of it as a tree T_0, consisting of 1 vertex and no edges. Then pick an increasing sequence of trees $T_0 \subset T_1 \subset T_2 \subset ... \subset T_e = T$ such that T_i has exactly i edges. We will prove the result in turn for T_0, T_1, T_2 and continue until we eventually prove it for T. We will call the ith edge that we add E_i, so $T_i = T_{i-1} \cup E_i$. Moreover, we will label the vertex where the

new edge gets added by t_i, so $V(E_i) \cap V(T_{i-1}) = \{t_i\}$ for $i \geq 1$. Note that a given vertex can receive several labels, one label or no labels at all; in fact, the number of labels a vertex has is always its degree minus 1 EXCEPT for the starting vertex, which has the extra label t_1.

This means that given a homomorphism $H \in \mathrm{Hom}(T_{i-1}, X)$, there are exactly $\deg(H(t_i))$ ways of extending H to some $H' \in \mathrm{Hom}(T_i, X)$. Moreover, if $H \in \mathrm{Hom}_{\mathcal{P}}(T_{i-1}, X)$, there are at least $\deg(H(t_i))\frac{q_{T_i}}{q_{T_{i-1}}} - p$ ways of extending H to some $H' \in \mathrm{Hom}_{\mathcal{P}}(T_i, X)$.

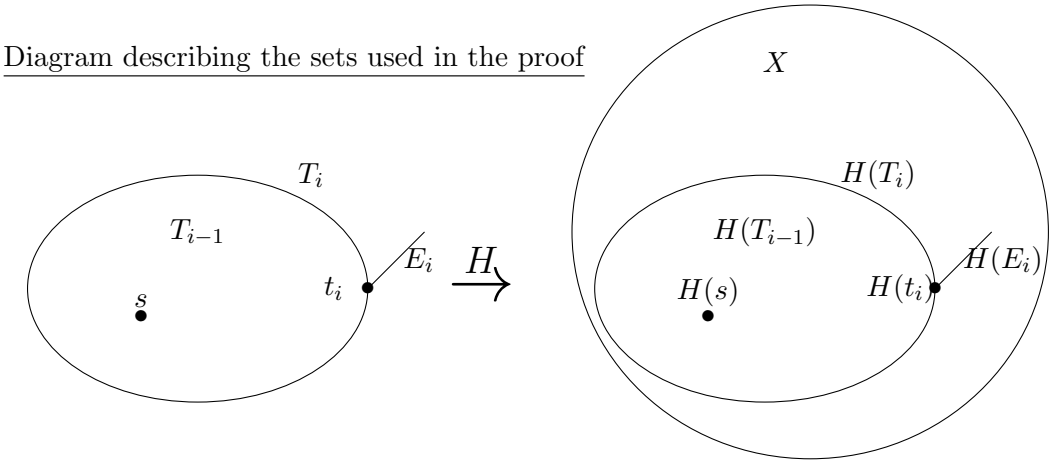
2.2 Proof of Theorem 1 (Szegedy)

We are going to define a probability distribution $\mu_i : \mathrm{Hom}(T_i, X) \to [0, 1]$, by induction on i. An element of $\mathrm{Hom}(T_1, X)$ is just a single directed edge. There are nd directed edges so there are nd elements of $\mathrm{Hom}(T_1, X)$. μ_1 is going to pick one uniformly at random, so $\mu_1(H) = \frac{1}{nd}$ for any $H \in \mathrm{Hom}(T_1, X)$.

For $i > 1$, pick $H \in \mathrm{Hom}(T_{i-1}, X)$. There are $\deg(H(t_i))$ possible choices for adding an extra edge to be the image of E_i. For every one of these, say H', we define $\mu_i(H') = \frac{\mu_{i-1}(H)}{\deg(H(t_i))}$. We are essentially just picking one extra edge to add to H_{i-1} uniformly at random amongst all the candidates.

Alternatively, $\mu_i(H') = \frac{1}{nd \cdot \prod_{j=2}^{i} \deg(H'(t_j))}$. This is also equal to $\frac{1}{nd \cdot \prod_{s \in T_i} \deg(H'(s))^{\deg(s)-1}}$ because every vertex $s \in T_i$ has exactly $\deg(s) - 1$ labels, when we exclude the extra label t_1. In particular, it doesn't depend on which order we added the edges to the tree, but only on what the final tree T_i looks like.

Given a vertex $x \in X$, we will define a probability distribution $\mu_0(x) = \frac{\deg(x)}{nd}$. This makes sense because under the identification of $\mathrm{Hom}(T_0, X)$ to X, the definition of μ_0 is consistent with the definition of μ_i for larger i and it satisfies the same properties.



2.2.1 Size of $\mathrm{Hom}(T_i, X)$

We claim that this probability distribution has the following property for any given vertex $x \in X$ and any vertex $s \in T_i$:

$$\sum_{H \in \mathrm{Hom}(T_i,X)\,:\,H(s)=x} \mu_i(H) = \mu_0(x). \tag{2.1}$$

Indeed, when $i = 1$, we have:

$$\sum_{H \in \mathrm{Hom}(T_1,X)\,:\,H(s)=x} \mu_1(H) = \sum_{F \in \mathrm{Hom}(E_1,X)\,:\,F(s)=x} \frac{1}{nd} = \frac{\deg(x)}{nd} = \mu_0(x).$$

For larger i, we have if $s \in T_{i-1}$:

$$\begin{aligned} \sum_{H' \in \mathrm{Hom}(T_i,X)\,:\,H'(s)=x} \mu_i(H') &= \sum_{\substack{H \in \mathrm{Hom}(T_{i-1},X)\,:\,H(s)=x \\ F \in Hom(E_i,X)\,:\,F(t_i)=H(t_i)}} \frac{\mu_{i-1}(H)}{\deg(H(t_i))} \\ &= \sum_{H \in \mathrm{Hom}(T_{i-1},X)\,:\,H(s)=x} \mu_{i-1}(H). \end{aligned}$$

and then this in turn is equal to $\mu_0(x)$ by the induction hypothesis (2.1).

If $s \notin T_{i-1}$, then $s \in E_i$, so we get:

$$\sum_{H' \in \mathrm{Hom}(T_i, X) \,:\, H'(s) = x} \mu_i(H') \;=\; \sum_{\substack{H \in \mathrm{Hom}(T_{i-1}, X) \\ F \in \mathrm{Hom}(E_i, X) \,:\, F(t_i) = H(t_i)\,,\, F(s) = x}} \frac{\mu_{i-1}(H}{\deg(H(t_i))}$$

$$= \sum_{F \in \mathrm{Hom}(E_i, X)\,,\, F(s) = x} \left(\sum_{H \in \mathrm{Hom}(T_{i-1}, X)\,,\, H(t_i) = F(t_i)} \frac{\mu_{i-1}(H)}{\deg(F(t_i))} \right).$$

Using the induction hypothesis (2.1) applied to $F(t_i)$ and t_i, we get that this is equal to:

$$\sum_{F \in \mathrm{Hom}(E_i, X)\,,\, F(s) = x} \left(\frac{\deg(F(t_i))}{nd} \cdot \frac{1}{\deg(F(t_i))} \right) = \sum_{F \in \mathrm{Hom}(E_i, X)\,,\, F(s) = x} \frac{1}{nd} = \frac{\deg(x)}{nd} = \mu_0(x).$$

So by induction, the claim is proved.

Consider the entropy $D(\mu_i) = \sum_{H \in \mathrm{Hom}(T_i, X)} - \ln(\mu_i(H))\mu_i(H)$. It is maximal when μ_i is uniform on $\mathrm{Hom}(T_i, X)$, therefore

$$D(\mu_i) \leq \sum_{H \in \mathrm{Hom}(T_i, X)} - \ln\left(\frac{1}{|\,\mathrm{Hom}(T_i, X)|} \right) \frac{1}{|\,\mathrm{Hom}(T_i, X)|} = \ln(|\,\mathrm{Hom}(T_i, X)|).$$

We now want to calculate $D(\mu_i)$ to get a lower bound on $|\,\mathrm{Hom}(T_i, X)|$. For $i = 1$, we have $\mu_1(H) = \frac{1}{nd}$ for all $H \in \mathrm{Hom}(T_1, X)$, so $D(\mu_1) = \sum_{H \in \mathrm{Hom}(T_1, X)} - \ln(\frac{1}{nd})\frac{1}{nd} = \ln(nd)$.

For larger i, we have, for every element $H' \in \mathrm{Hom}(T_i, X)$, $H' = H \cup F$, where H is an element of $\mathrm{Hom}(T_{i-1}, X)$ and F is an element of $\mathrm{Hom}(E_i, X)$. Using the definition of μ_i, we can rewrite $D(\mu_i)$ as:

$$
\begin{aligned}
D(\mu_i) &= \sum_{\substack{H\in\mathrm{Hom}(T_{i-1},X)\\ F\in\mathrm{Hom}(E_i,X)\\ F(t_i)=H(t_i)}} -\ln\left(\frac{\mu_{i-1}(H)}{\deg(H(t_i))}\right)\frac{\mu_{i-1}(H)}{\deg(H(t_i))} \\
&= \sum_{\substack{H\in\mathrm{Hom}(T_{i-1},X)\\ F\in\mathrm{Hom}(E_i,X)\\ F(t_i)=H(t_i)}} \left[-\ln(\mu_{i-1}(H))\frac{\mu_{i-1}(H)}{\deg(H(t_i))}\right] \\
&\quad + \sum_{\substack{H\in\mathrm{Hom}(T_{i-1},X)\\ F\in\mathrm{Hom}(E_i,X)\\ F(t_i)=H(t_i)}} \left[\ln(\deg(H(t_i)))\frac{\mu_{i-1}(H)}{\deg(H(t_i))}\right] \\
&= \sum_{H\in\mathrm{Hom}(T_{i-1},X)} [-\ln(\mu_{i-1}(H))\mu_{i-1}(H)] \\
&\quad + \sum_{\substack{F\in\mathrm{Hom}(E_i,X)\\ H\in\mathrm{Hom}(T_{i-1},X)\\ H(t_i)=F(t_i)}} \left[\ln(\deg(F(t_i)))\frac{\mu_{i-1}(H)}{\deg(F(t_i))}\right].
\end{aligned}
$$

Using property (2.1) applied to $F(t_i)$ and H in the second term, we can rewrite this as:

$$
\begin{aligned}
D(\mu_i) &= \sum_{H\in\mathrm{Hom}(T_{i-1},X)} [-\ln(\mu_{i-1}(H))\mu_{i-1}(H)] \\
&\quad + \sum_{F\in\mathrm{Hom}(E_i,X)} \left[\ln(\deg(F(t_i)))\frac{\deg(F(t_i))/nd}{\deg(F(t_i))}\right] \\
&= D(\mu_{i-1}) + \sum_{x\in X} |\{F\in\mathrm{Hom}(E_i,X)\,:\,F(t_i)=x\}| \left[\ln(\deg(x))\frac{1}{nd}\right] \\
&= D(\mu_{i-1}) + \sum_{x\in X}\left[\ln(\deg(x))\frac{\deg(x)}{nd}\right] \\
&= D(\mu_{i-1}) + \sum_{x\in X}\left[\ln\left(\frac{\deg(x)}{nd}\right)\frac{\deg(x)}{nd}\right] + \sum_{x\in X}\left[\ln(nd)\frac{\deg(x)}{nd}\right] \\
&= D(\mu_{i-1}) - D(\mu_0) + \ln(nd).
\end{aligned}
$$

Now we use the entropy inequality on μ_0 to say that $D(\mu_0)\leq\ln(n)$ and this gives us:

$$D(\mu_i)\geq D(\mu_{i-1})+\ln(d).$$

So by induction, we have $D(\mu_i) \geq \ln(nd) + (i-1)\ln(d) = \ln\left(nd^i\right)$ and therefore we have at least nd^i homomorphisms from T_i to X. This proves Theorem 1.

Also notice that this is a stability result: if we are close to equality in Theorem 1, that implies we need to be close to equality in the entropy inequality for μ_0, which implies that μ_0 should be close to a uniform distribution. But $\mu_0(x) = \frac{\deg(x)}{nd}$, so we would need all the degrees to be close to equal. We have equality in Theorem 1 if and only if the graph is regular, and we are close to equality only when the graph is close to regular.

2.3 Proof of Theorem 2

2.3.1 Probability of having property $\mathcal{P}$

Given a random element H' of $\mathrm{Hom}(T_i, X)$ (chosen according to the probability distribution μ_i), we want to know the probability it has property $\mathcal{P}$. We'll call this probability $\mathbb{P}_i$. By our assumption on $\mathcal{P}$, given an element H of $\mathrm{Hom}(T_{i-1}, X)$, there are $\geq \deg(H(t_i))\frac{q_{T_i}}{q_{T_{i-1}}} - p$ ways to extend it.

With that said, the probability of a homomorphism $H' \in \mathrm{Hom}(T_i, X)$ having property $\mathcal{P}$ is at least:

$$\begin{aligned}
\mathbb{P}_i &\geq \sum_{H\in \mathrm{Hom}_{\mathcal{P}}(T_{i-1},X)} \frac{\mu_{i-1}(H)}{\deg(H(t_i))}\left[\deg(H(t_i))\frac{q_{T_i}}{q_{T_{i-1}}} - p\right] \\
&= \frac{q_{T_i}}{q_{T_{i-1}}}\left[\sum_{H\in \mathrm{Hom}_{\mathcal{P}}(T_{i-1},X)} \mu_{i-1}(H)\right] - p\left[\sum_{H\in \mathrm{Hom}_{\mathcal{P}}(T_{i-1},X)} \frac{\mu_{i-1}(H)}{\deg(H(t_i))}\right] \\
&\geq \frac{q_{T_i}}{q_{T_{i-1}}}\mathbb{P}_{i-1} - p\left[\sum_{H\in \mathrm{Hom}(T_{i-1},X)} \frac{\mu_{i-1}(H)}{\deg(H(t_i))}\right] \\
&= \frac{q_{T_i}}{q_{T_{i-1}}}\mathbb{P}_{i-1} - p\left[\sum_{\substack{x\in X \\ H\in \mathrm{Hom}(T_{i-1},X) \\ H(t_i)=x}} \frac{\mu_{i-1}(H)}{\deg(x)}\right] \\
&= \frac{q_{T_i}}{q_{T_{i-1}}}\mathbb{P}_{i-1} - p\left[\sum_{x\in X} \frac{\deg(x)}{dn}\frac{1}{\deg(x)}\right] \\
&= \frac{q_{T_i}}{q_{T_{i-1}}}\mathbb{P}_{i-1} - p\cdot\frac{1}{d}.
\end{aligned}$$

(We use property (2.1) again to go from the fourth line to the fifth.)
We claim that $\mathbb{P}_i \geq q_{T_i} - \left[\sum_{j=1}^{i}\frac{q_{T_i}}{q_{T_j}}\right]\frac{p}{d}$. When $i=0$, we have $\mathbb{P}_0 = 1$ and $q_{T_0} = 1$ by our assumptions on $\mathcal{P}$, which agrees with the formula. For larger i, we use induction and have, from the above inequality,

$$\begin{aligned}
\mathbb{P}_i &\geq \frac{q_{T_i}}{q_{T_{i-1}}}\cdot\left[q_{T_{i-1}} - \left[\sum_{j=1}^{i-1}\frac{q_{T_{i-1}}}{q_{T_j}}\right]\right]\frac{p}{d} - \frac{p}{d} \\
&= q_{T_i} - \left[\sum_{j=1}^{i-1}\frac{q_{T_i}}{q_{T_j}}\right]\frac{p}{d} - \frac{p}{d} \\
&= q_{T_i} - \left[\sum_{j=1}^{i}\frac{q_{T_i}}{q_{T_j}}\right]\frac{p}{d}.
\end{aligned}$$

Thus,

$$\mathbb{P}_i \geq q_{T_i} - i\frac{p}{d}. \tag{2.2}$$

We can also come up with an upper bound using similar methods:

$$\begin{aligned}
\mathbb{P}_i &\leq \sum_{H \in \mathrm{Hom}_{\mathcal{P}}(T_{i-1},X)} \frac{\mu_{i-1}(H)}{\deg(H(t_i))} \deg(H(t_i)) \frac{q_{T_i}}{q_{T_{i-1}}} \\
&= \frac{q_{T_i}}{q_{T_{i-1}}} \sum_{H \in \mathrm{Hom}_{\mathcal{P}}(T_{i-1},X)} \mu_{i-1}(H) \\
&= \frac{q_{T_i}}{q_{T_{i-1}}} \mathbb{P}_{i-1} \\
&= \ldots \\
&= \frac{q_{T_i}}{q_{T_0}} \mathbb{P}_0 \\
&= q_{T_i}
\end{aligned}$$

Therefore

$$q_{T_i} \geq \mathbb{P}_i \geq q_{T_i} - i\frac{p}{d}. \tag{2.3}$$

2.3.2 Size of $\mathrm{Hom}_{\mathcal{P}}(T_i, X)$

Step 1: Setting up a proof by induction

So we know the number of homomorphisms, and we know the probability that one of them satisfies property $\mathcal{P}$. From this, we want to find the number of homomorphisms that satisfy $\mathcal{P}$. This is slightly more complicated than it seems because μ_i is not uniform. However, we can still find a lower bound. First, we will use the entropy inequality for the induced probability distribution on $\mathrm{Hom}(T_i, X)$:

$$\begin{aligned}
\ln(|\mathrm{Hom}_{\mathcal{P}}(T_i,X)|) &= \sum_{H' \in \mathrm{Hom}_{\mathcal{P}}(T_i,X)} -\ln\left(\frac{1}{|\mathrm{Hom}_{\mathcal{P}}(T_i,X)|}\right) \frac{1}{|\mathrm{Hom}_{\mathcal{P}}(T_i,X)|} \\
&\geq \sum_{H' \in \mathrm{Hom}_{\mathcal{P}}(T_i,X)} -\ln\left(\frac{\mu_i(H)}{\mathbb{P}_i}\right) \frac{\mu_i(H)}{\mathbb{P}_i}
\end{aligned}$$

$$\ln(|\mathrm{Hom}_{\mathcal{P}}(T_i,X)|) \geq \ln(\mathbb{P}_i) + \frac{\sum_{H' \in \mathrm{Hom}_{\mathcal{P}}(T_i,X)} -\ln(\mu_i(H'))\mu_i(H')}{\mathbb{P}_i}. \tag{2.4}$$

So now we want to find a lower bound on $\sum_{H' \in \mathrm{Hom}_{\mathcal{P}}(T_i,X)} -\ln(\mu_i(H'))\mu_i(H')$.

Claim:

$$\sum_{H' \in \mathrm{Hom}_{\mathcal{P}}(T_i, X)} -\ln(\mu_i(H'))\mu_i(H') \geq q_{T_i} \cdot \ln\left(nd^i\right) - i(i+1)\frac{\ln(n)p}{d} \tag{2.5}$$

We will prove (2.5) by induction on i. For $i = 1$, we get:

$$\sum_{H' \in \mathrm{Hom}_{\mathcal{P}}(T_1, X)} -\ln(\mu_1(H'))\mu_1(H') = \sum_{H' \in \mathrm{Hom}_{\mathcal{P}}(T_1, X)} -\ln\left(\frac{1}{nd}\right)\frac{1}{nd} \geq$$

$$\frac{\ln(nd)}{nd} \sum_{H \in \mathrm{Hom}_{\mathcal{P}}(T_0, X)} \left[\deg(H(t_1))\frac{q_{T_1}}{q_{T_0}} - p\right] = \frac{\ln(nd)}{nd} \sum_{x \in X} [\deg(x) q_{T_1} - p] =$$

$$\frac{\ln(nd)}{nd}[ndq_{T_1} - np] \geq q_{T_1}\ln(nd) - \frac{\ln(n^2)p}{d} = q_{T_1}\ln(nd) - 2\frac{\ln(n)p}{d}$$

so it is true for $i = 1$.

Step 2: For larger i, expressing $\sum_{H' \in \mathrm{Hom}_{\mathcal{P}}(T_i, X)} -\ln(\mu_i(H'))\mu_i(H')$ as a linear combination of three terms

We proceed as follows:

$$\begin{aligned}
&\sum_{H' \in \mathrm{Hom}_{\mathcal{P}}(T_i, X)} -\ln(\mu_i(H'))\mu_i(H') \\
= &\sum_{\substack{x \in X \\ H \in \mathrm{Hom}_{\mathcal{P}}(T_{i-1}, X) : H(t_i) = x \\ F \in \mathrm{Hom}(E_i, X) : F(t_i) = x \\ H \cup F \text{ has } \mathcal{P}}} -\ln\left(\frac{\mu_{i-1}(H)}{\deg(x)}\right)\frac{\mu_{i-1}(H)}{\deg(x)} \\
\geq &\sum_{\substack{x \in X \\ H \in \mathrm{Hom}_{\mathcal{P}}(T_{i-1}, X) \\ H(t_i) = x}} -\ln\left(\frac{\mu_{i-1}(H)}{\deg(x)}\right)\frac{\mu_{i-1}(H)}{\deg(x)}\left(\deg(x)\frac{q_{T_i}}{q_{T_{i-1}}} - p\right)
\end{aligned}$$

$$
\begin{aligned}
&= \frac{q_{T_i}}{q_{T_{i-1}}} \left[\sum_{\substack{x \in X \\ H \in \mathrm{Hom}_{\mathcal{P}}(T_{i-1}, X) \\ H(t_i) = x}} - \ln\left(\frac{\mu_{i-1}(H)}{\deg(x)}\right) \mu_{i-1}(H) \right] \\
&\quad -p \left[\sum_{\substack{x \in X \\ H \in \mathrm{Hom}_{\mathcal{P}}(T_{i-1}, X) \\ H(t_i) = x}} - \ln\left(\frac{\mu_{i-1}(H)}{\deg(x)}\right) \frac{\mu_{i-1}(H)}{\deg(x)} \right] \\
&= \frac{q_{T_i}}{q_{T_{i-1}}} \left[\sum_{\substack{x \in X \\ H \in \mathrm{Hom}_{\mathcal{P}}(T_{i-1}, X) \\ H(t_i) = x}} - \ln(\mu_{i-1}(H)) \mu_{i-1}(H) \right] \\
&\quad + \frac{q_{T_i}}{q_{T_{i-1}}} \left[\sum_{\substack{x \in X \\ H \in \mathrm{Hom}_{\mathcal{P}}(T_{i-1}, X) \\ H(t_i) = x}} \ln\left(\deg(x)\right) \mu_{i-1}(H) \right] \\
&\quad -p \left[\sum_{\substack{x \in X \\ H \in \mathrm{Hom}_{\mathcal{P}}(T_{i-1}, X) \\ H(t_i) = x}} - \ln\left(\frac{\mu_{i-1}(H)}{\deg(x)}\right) \frac{\mu_{i-1}(H)}{\deg(x)} \right].
\end{aligned}
$$

This is a linear combination of three terms; we will simplify each of these terms separately.

The first term

$$
\sum_{\substack{x \in X \\ H \in \mathrm{Hom}_{\mathcal{P}}(T_{i-1}, X) \\ H(t_i) = x}} - \ln(\mu_{i-1}(H)) \mu_{i-1}(H) = \sum_{H \in \mathrm{Hom}_{\mathcal{P}}(T_{i-1}, X)} - \ln(\mu_{i-1}(H)) \mu_{i-1}(H).
$$

And now by the induction hypothesis, we get:

$$\sum_{\substack{x\in X\\ H\in \mathrm{Hom}_{\mathcal{P}}(T_{i-1},X)\\ H(t_i)=x}} -\ln(\mu_{i-1}(H))\mu_{i-1}(H) \geq q_{T_{i-1}}\cdot \ln\left(nd^{i-1}\right) - i(i-1)\frac{\ln(n)p}{d}.$$

The second term

First of all, let $s \in v(T_i)$ be arbitrary. Similarly to before, we pick a sequence of trees $\{s\} = T'_0 \subset T'_1 \subset T'_2 \subset ... \subset T'_i = T_i$, where T'_j has exactly j vertices. We will call the jth edge that we add E'_j, so $T'_j = T'_{j-1} \cup E'_j$. Moreover, we will label the vertex where the new edge gets added by t'_j, so $v(E'_j) \cap v(T'_{j-1}) = \{t'_j\}$.

The quantity we will be looking at is $\sum_{H'\in \mathrm{Hom}_{\mathcal{P}}(T'_j,X)} \ln\left(\deg(H'(s))\right)\mu_j(H')$, which matches the second term when $j = i-1$ and $s = t_i$. We will proceed by induction on j:

$$
\begin{aligned}
& \sum_{\substack{x\in X \\ H'\in \mathrm{Hom}_{\mathcal{P}}(T'_j,X) \\ H(s)=x}} \ln\left(\deg(x)\right)\mu_j(H') \\
= & \sum_{\substack{x\in X \\ H\in \mathrm{Hom}_{\mathcal{P}}(T'_{j-1},X) \\ H(s)=x \\ F\in \mathrm{Hom}(E'_j,X) \\ F(t'_j)=H(t'_j) \\ F\cup H \text{ has } \mathcal{P}}} \ln(\deg(x))\frac{\mu_{j-1}(H)}{\deg(H(t'_j))} \\
\geq & \sum_{\substack{x\in X \\ H\in \mathrm{Hom}_{\mathcal{P}}(T'_{j-1},X) \\ H(s)=x}} \ln(\deg(x))\mu_{j-1}(H)\frac{\frac{q_{T'_j}}{q_{T'_{j-1}}}\deg(H(t'_j))-p}{\deg(H(t'_j))} \\
= & \frac{q_{T'_j}}{q_{T'_{j-1}}}\left[\sum_{\substack{x\in X \\ H\in \mathrm{Hom}_{\mathcal{P}}(T'_{j-1},X) \\ H(s)=x}} \ln(\deg(x))\mu_{j-1}(H)\right] \\
& -p\left[\sum_{\substack{x\in X \\ H\in \mathrm{Hom}_{\mathcal{P}}(T'_{j-1},X) \\ H(s)=x}} \ln(\deg(x))\frac{\mu_{j-1}(H)}{\deg(H(t'_j))}\right] \\
\geq & \frac{q_{T'_j}}{q_{T'_{j-1}}}\left[\sum_{\substack{x\in X \\ H\in \mathrm{Hom}_{\mathcal{P}}(T'_{j-1},X) \\ H(s)=x}} \ln(\deg(x))\mu_{j-1}(H)\right] - p\left[\sum_{H\in \mathrm{Hom}(T'_{j-1},X)} \ln(n)\frac{\mu_{j-1}(H)}{\deg(H(t'_j))}\right]
\end{aligned}
$$

$$
\begin{aligned}
&= \frac{q_{T'_j}}{q_{T'_{j-1}}} \left[\sum_{\substack{x\in X \\ H\in \mathrm{Hom}_{\mathcal{P}}(T'_{j-1},X) \\ H(s)=x}} \ln(\deg(x))\mu_{j-1}(H) \right] - p \left[\sum_{\substack{y\in X \\ H\in \mathrm{Hom}(T'_{j-1},X) \\ H(t'_j)=y}} \ln(n)\frac{\mu_{j-1}(H)}{\deg(y)} \right] \\
&= \frac{q_{T'_j}}{q_{T'_{j-1}}} \left[\sum_{\substack{x\in X \\ H\in \mathrm{Hom}_{\mathcal{P}}(T'_{j-1},X) \\ H(s)=x}} \ln(\deg(x))\mu_{j-1}(H) \right] - p \left[\sum_{y\in X} \ln(n)\frac{\deg(y)}{\deg(y)\cdot nd} \right] \\
&= \frac{q_{T'_j}}{q_{T'_{j-1}}} \left[\sum_{\substack{x\in X \\ H\in \mathrm{Hom}_{\mathcal{P}}(T'_{j-1},X) \\ H(s)=x}} \ln(\deg(x))\mu_{j-1}(H) \right] - p\frac{\ln(n)}{d}.
\end{aligned}
$$

Note that we used property (2.1) to go from the 6th line to the 7th line. Once we have done this calculation for j going from $i-1$ to 1, we end up with:

$$
\begin{aligned}
&\sum_{H'\in \mathrm{Hom}_{\mathcal{P}}(T_{i-1},X)} \ln\left(\deg(H'(s))\right)\mu_{i-1}(H') \\
&\geq \prod_{j=1}^{i-1} \frac{q_{T'_j}}{q_{T'_{j-1}}} \cdot \left[\sum_{\substack{x\in X \\ H\in \mathrm{Hom}_{\mathcal{P}}(T'_0,X) \\ H(s)=x}} \ln(\deg(x))\mu_0(H) \right] - \ln(n) \left[\sum_{j=1}^{i-1} \prod_{l=j+1}^{i-1} \frac{q_{T'_j}}{q_{T'_{j-1}}} \right] \frac{p}{d} \\
&\geq q_{T_{i-1}} \left[\sum_{\substack{x\in X \\ H\in \mathrm{Hom}_{\mathcal{P}}(T'_0,X) \\ H(s)=x}} \ln(\deg(x))\mu_0(H) \right] - (i-1)\frac{\ln(n)p}{d}
\end{aligned}
$$

$$
\begin{aligned}
&= q_{T_{i-1}} \left[\sum_{x \in X} \ln(\deg(x)) \frac{\deg(x)}{dn} \right] - (i-1)\frac{\ln(n)p}{d} \\
&= q_{T_{i-1}} \left[\sum_{x \in X} \ln\left(\frac{\deg(x)}{nd}\right) \frac{\deg(x)}{dn} + \ln(nd)\frac{\deg(x)}{dn} \right] - (i-1)\frac{\ln(n)p}{d} \\
&= q_{T_{i-1}} \cdot [-D(\mu_0) + \ln(nd)] - (i-1)\frac{\ln(n)p}{d}.
\end{aligned}
$$

Using the entropy inequality on μ_0, we get that $D(\mu_0) \leq \ln(n)$ and so therefore this is at least:

$$q_{T_{i-1}} \ln(d) - (i-1)\frac{\ln(n)p}{d}.$$

The third term

The third term is:

$$
\begin{aligned}
&\sum_{\substack{x \in X \\ H \in \mathrm{Hom}_{\mathcal{P}}(T_{i-1}, X) \\ H(t_i) = x}} -\ln\left(\frac{\mu_{i-1}(H)}{\deg(x)}\right)\frac{\mu_{i-1}(H)}{\deg(x)} \\
\leq &\sum_{\substack{x \in X \\ H \in \mathrm{Hom}(T_{i-1}, X) \\ H(t_i) = x}} -\ln\left(\frac{\mu_{i-1}(H)}{\deg(x)}\right)\frac{\mu_{i-1}(H)}{\deg(x)} \\
= &\sum_{\substack{x \in X \\ H \in \mathrm{Hom}(T_{i-1}, X) \\ H(t_i) = x}} -\ln\left(\frac{\mu_{i-1}(H)}{\deg(x)}\right)\frac{\mu_{i-1}(H)\cdot d}{\deg(x)} + \sum_{\substack{x \in X \\ H \in \mathrm{Hom}(T_{i-1}, X) \\ H(t_i) = x}} \ln(d)\frac{\mu_{i-1}(H)}{\deg(x)} \\
= &\frac{1}{d}\sum_{\substack{x \in X \\ H \in \mathrm{Hom}(T_{i-1}, X) \\ H(t_i) = x}} -\ln\left(\frac{\mu_{i-1}(H)\cdot d}{\deg(x)}\right)\frac{\mu_{i-1}(H)\cdot d}{\deg(x)} + \sum_{x \in X} \ln(d)\frac{\deg(x)/nd}{\deg(x)}.
\end{aligned}
$$

Note that because of property 2.1, $\sum_{x \in X\,;\, H \in \mathrm{Hom}(T_{i-1}, X)\,:\, H(t_i) = x} \frac{\mu_{i-1}(H)\cdot d}{\deg(x)} = \sum_{x \in X} \frac{\deg(x)}{nd}\frac{d}{\deg(x)} = \sum_{x \in X} \frac{1}{n} = 1$ so $\frac{\mu_{i-1}(H)\cdot d}{\deg(x)}$ is a probability distribution on $\mathrm{Hom}(T_{i-1}, X)$. That means we can do another entropy inequality and get that this is less than or equal to:

$$\begin{aligned}
& \sum_{\substack{x\in X\\ H\in \mathrm{Hom}_{\mathcal{P}}(T_{i-1},X)\\ H(t_i)=x}} -\ln\left(\frac{\mu_{i-1}(H)}{\deg(x)}\right)\frac{\mu_{i-1}(H)}{\deg(x)} \\
\leq\ & \frac{1}{d}\sum_{\substack{x\in X\\ H\in \mathrm{Hom}(T_{i-1},X)\\ H(t_i)=x}} -\ln\left(\frac{1}{|\mathrm{Hom}(T_{i-1},X)|}\right)\frac{\mu_{i-1}(H)\cdot d}{\deg(x)} + \sum_{x\in X}\ln(d)\frac{1}{nd} \\
=\ & \frac{1}{d}\ln(|\mathrm{Hom}(T_{i-1},X)|) + \frac{\ln(d)}{d}.
\end{aligned}$$

We now need to bound $|\mathrm{Hom}(T_{i-1},X)|$ from above. We don't need to do anything fancy for this: we just pick i vertices, and sometimes they will form a copy of T_{i-1}. This implies that $|\mathrm{Hom}(T_{i-1},X)| \leq n^i$. Plugging this back into the formula gives:

$$\sum_{\substack{x\in X\\ H\in \mathrm{Hom}_{\mathcal{P}}(T_{i-1},X)\\ H(t_i)=x}} -\ln\left(\frac{\mu_{i-1}(H)}{\deg(x)}\right)\frac{\mu_{i-1}(H)}{\deg(x)} \leq \frac{\ln(n^i d)}{d} \leq (i+1)\frac{\ln(n)}{d}.$$

Putting the three terms back together again

Thus, by adding up the three terms back together again, we get:

$$\begin{aligned}
& \sum_{\substack{x\in X\\ H\in \mathrm{Hom}_{\mathcal{P}}(T_{i-1},X):H(t_i)=x\\ F\in \mathrm{Hom}(E_i,X):F(t_i)=x\\ H\cup F \text{ has } \mathcal{P}}} -\ln\left(\frac{\mu_{i-1}(H)}{\deg(x)}\right)\frac{\mu_{i-1}(H)}{\deg(x)} \\
\geq\ & \frac{q_{T_i}}{q_{T_{i-1}}}\left[q_{T_{i-1}}\ln\left(nd^{i-1}\right) - i(i-1)\frac{\ln(n)p}{d}\right] + \frac{q_{T_i}}{q_{T_{i-1}}}\left[q_{T_{i-1}}\ln(d) - (i-1)\frac{\ln(n)p}{d}\right] \\
& -p\left[(i+1)\frac{\ln(n)}{d}\right] \\
\geq\ & q_{T_i}\left[\ln(nd^{i-1}) + \ln(d)\right] - \frac{\ln(n)p}{d}\frac{q_{T_i}}{q_{T_{i-1}}}\left[i(i-1) + (i-1) + (i+1)\right] \\
=\ & q_{T_i}\ln(nd^i) - \frac{\ln(n)p}{d}i(i+1).
\end{aligned}$$

This completes the proof of proposition (2.5) by induction.

Finishing up the lower bound on $|M_i|$

Now we can combine (2.5) and (2.4), to get:

$$\ln(|\operatorname{Hom}_{\mathcal{P}}(T_i, X)|) \geq \ln(\mathbb{P}_i) + \frac{q_{T_i}}{\mathbb{P}_i} \ln\left(nd^i\right) - i(i+1)\frac{\ln(n)p}{d}.$$

Remember from (2.3) that $q_{T_i} - i\frac{p}{d} \leq \mathbb{P}_i \leq q_{T_i}$, so we get: $\ln(|\operatorname{Hom}_{\mathcal{P}}(T_i, X)|) \geq \ln\left(q_{T_i} - i\frac{p}{d}\right) + \frac{q_{T_i}}{q_{T_i}} \ln\left(nd^i\right) - i(i+1)\frac{\ln(n)p}{d}$. Using the concavity of the ln function close to 1, we get:

$$\begin{aligned}
\ln(|\operatorname{Hom}_{\mathcal{P}}(T_i, X)|) &\geq \ln\left(q_{T_i} - i\frac{p}{d}\right) + \frac{q_{T_i}}{q_{T_i}} \ln\left(nd^i\right) - i(i+1)\frac{\ln(n)p}{d} \\
&\geq \ln\left(q_{T_i} - i\frac{p}{d}\right) + \ln\left(nd^i\right) + \ln\left(1 - i(i+1)\frac{\ln(n)p}{d}\right) \\
&= \ln\left(\left(q_{T_i} - i\frac{p}{d}\right)\left(nd^i\right)\left(1 - i(i+1)\frac{\ln(n)p}{d}\right)\right) \\
&\geq \ln\left(q_{T_i} \cdot nd^i \cdot \left(1 - i(i+1)\frac{\ln(n)p}{d} - \frac{q_{T_i}^{-1} ip}{d}\right)\right) \\
&\geq \ln\left(q_{T_i} \cdot nd^i \cdot \left(1 - i(i+1+q_{T_i}^{-1})\frac{\ln(n)p}{d}\right)\right).
\end{aligned}$$

Thus:

$$|\operatorname{Hom}_{\mathcal{P}}(T_i, X)| \geq q_{T_i} nd^i \cdot \left(1 - i(i+1+q_{T_i}^{-1})\frac{\ln(n)p}{d}\right).$$

Setting $i = e$, this is exactly the statement of Theorem 2.

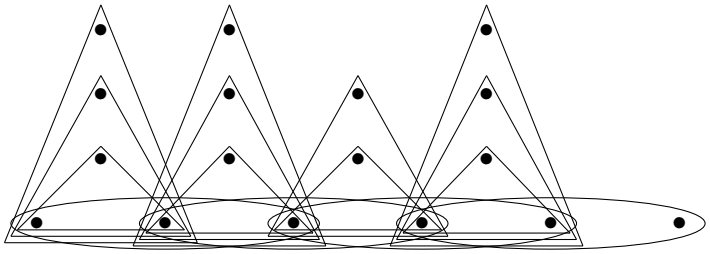
□

2.4 Lower bound on the number of copies of a tight k-hypertrees in a k-hypergraph

In this section, we will overview an important corollary of Theorem 2, where given some tight k-hypertree, we will find a lower bound on the number of copies of this hypertree in some larger k-hypergraph. But first, let us define what we mean by a tight k-hypertree:

Definition 11. *Given an integer k, a tight k-hypertree is a k-hypergraph whose edges can be labelled $E_1, E_2, E_3, E_4, ...E_e$ such that for every edge E_i, $i > 1$, there exists some $j(i) < i$ such that $|E_i \cap E_{j(i)}| = k - 1$ but $E_i \backslash E_{j(i)}$ does not intersect any E_l for $l < i$. Moreover, if $j(i) = j(i')$ for some i, i', then $E_i \cap E_{j(i)} = E_{i'} \cap E_{j(i')}$.*

Remark: a tight k-hypertree with e edges has $e + k - 1$ vertices.



Example of a tight 3-hypertree T

The corollary we will prove in this section is as follows:

Corollary 1. *For a tight k-hypertree $\mathcal{T}$ with e edges, and a larger k-hypergraph G with n vertices such that the average degree of a $k-1$-set in X is d, then there are at least $\binom{n}{k-1}(k-1)! \cdot (d/(k-1))^e \cdot \left(1 - O\left(\frac{\ln(n)}{d}\right)\right)$ copies of $\mathcal{T}$ inside G.*

To prove this, we first need a definition:

Definition 12. *An ordered l-set x is a set of size l equipped with a bijection $f_x : x \to \{1, 2, 3, ..., l\}$. We call this bijection the ordering of x.*

For every set of size l, there are exactly $l!$ orderings of it.

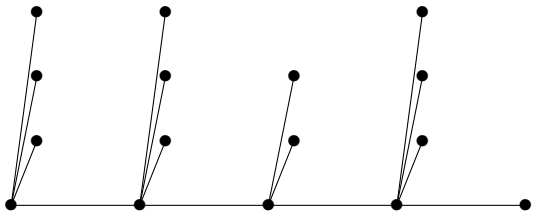
Proof of Corollary 1

We will define the graph X to have as its vertex set the collection of all ordered $(k-1)$-sets of vertices of G. This means that X has $\binom{n}{k-1}(k-1)!$ vertices.

The edges of X are defined to be the pairs of ordered $(k-1)$-sets , (x, y) , such that $x \cup y$ is an edge of G, and such that the two orderings agree on $x \cap y$, in other words, for all $z \in x \cap y$, $f_x(z) = f_y(z)$.

We'll also define the function θ that sends graphs in X to the corresponding k-hypergraphs in G, by sending each vertex to the corresponding $(k-1)$-sets and each edge to the corresponding edge in G. Similarly, we'll also define θ' that sends a tree T to the corresponding

tight k-hypertree $\mathcal{T}$ in the exact same way.

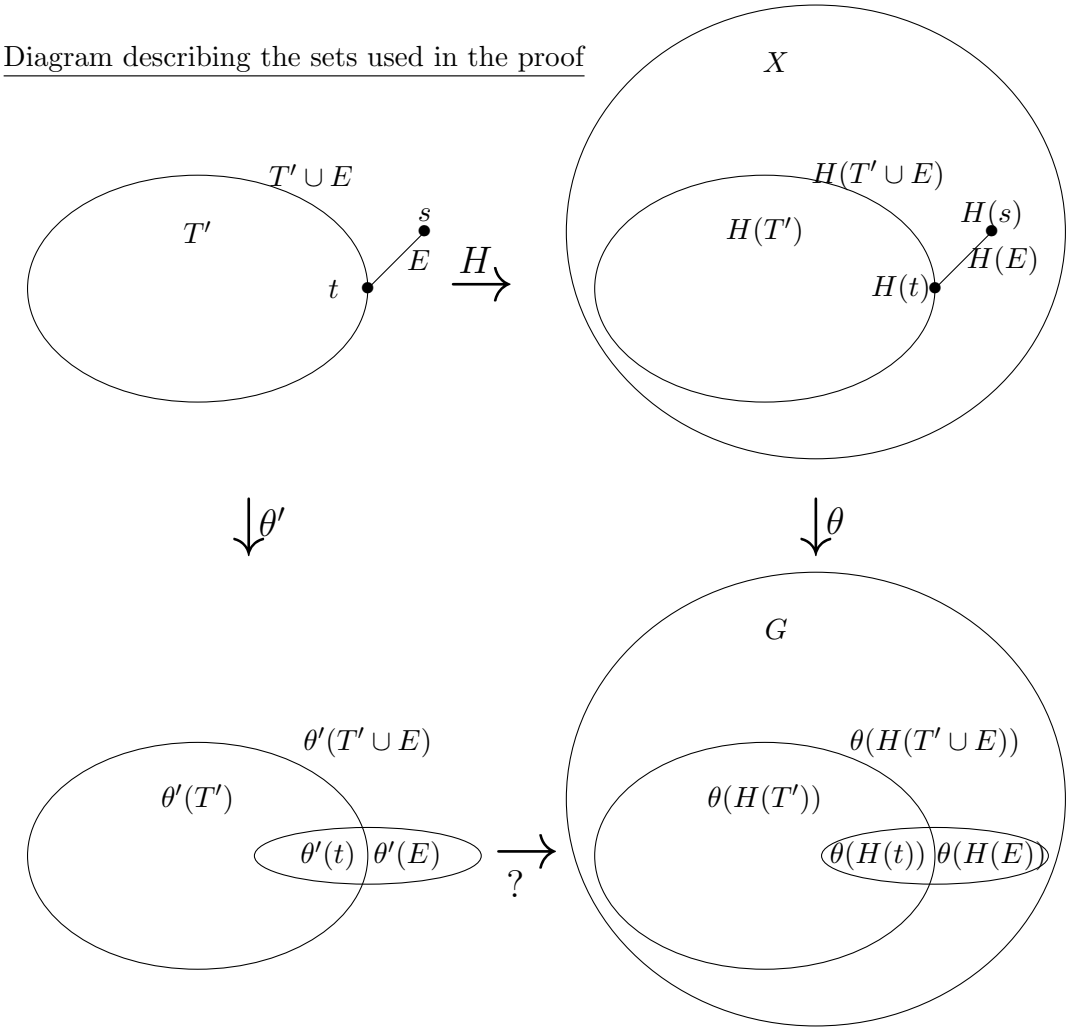


A graph T such that $\theta'(T) = \mathcal{T}$.

Given a particular tight k-hypertree $\mathcal{T}$ with e edges, why can we always find some tree T with $\theta'(T) = \mathcal{T}$? That is to say, a tree whose vertices are $k-1$-sets and whose edges represent k-sets that contain both its incident vertices?
Well, for every edge E_i of $\mathcal{T}$, we'll have exactly one corresponding edge E'_i in T. What are its incident vertices? For every $i > 1$, define $E_i \cap E_{j(i)}$ to be one of them. If there exists some i' with $j(i') = i$, then define $E_{i'} \cap E_i$ to be the other one. Notice that if i'' also has $j(i'') = i$, then this doesn't change what this second vertex is. After this process, we are left with a collection of edges, some of which have two incident vertices, some of which still only have one. For each edge that still only has one incident vertex, just pick a second incident vertex (different from the first) arbitrarily. We have thus constructed a tree T' with $\theta'(T) = \mathcal{T}$. Also notice that T has the same number of edges as $\mathcal{T}$.

We remove one leaf at a time from T to get a sequence of trees $T_1 \subset T_2 \subset ... \subset T_e = T$ such that T_i has exactly i edges. There is a corresponding sequence $\theta'(T_1) \subset \theta'(T_2) \subset \theta'(T_3) \subset ... \subset \theta'(T_e)$ such that $\theta'(T_i)$ has exactly i edges.

To use Theorem 2, we need some injective-like property $\mathcal{P}$ of homomorphisms. We will say that $H \in \mathrm{Hom}(T_i, G)$ has $\mathcal{P}$ if and only if the image of $H \circ \theta$ is a copy of $\theta'(T_i)$. In particular, $H \in \mathrm{Hom}(T, G)$ has $\mathcal{P}$ if and only if the image of $H \circ \theta$ is a copy of $\mathcal{T}$. On the diagram below, this is the equivalent of saying that there is an injective homomorphism where the question mark is, and this homomorphism makes the diagram commute.



We need to check that $\mathcal{P}$ is strongly injective-like. The first thing we need is to define some constant p, which we'll pick to be $p = e - 1$. For every subtree T' of T, we also need some $q_{T'}$; we will pick $q_{T'} = (k-1)^{-e(T')}$. Note that when T' consists of just a single vertex, $q_{T'} = 1$, so the second bullet point of Definition 9 (weakly injective-like) is immediately satisfied.

The second thing we need to check is that every homomorphism H from the single vertex T_0 to X satisfies $\mathcal{P}$. But $\theta'(T_0)$ is a collection of $k-1$ distinct points, while H picks out a single vertex of X. All the vertices in X are ordered $(k-1)$-sets of vertices in G so $\theta(H(T_0))$ is indeed a collection of $k-1$ distinct points and hence a copy of $\theta'(T_0)$. So the first bullet point of Definition 9 does hold.

To check the last bullet point, we are given a subtree $T' \subset T$, an edge E incident to T' at t, and homomorphism $H \in \mathrm{Hom}_{\mathcal{P}}(T', X)$. We are asked to count the number of homomorphisms $H' \in \mathrm{Hom}_{\mathcal{P}}(T' \cup E, X)$ that extend it. Because H has $\mathcal{P}$, we already know that $H \circ \theta$ is a copy of $\theta'(T')$. We want to add the image of E to it, which should be an edge of X that is connected to $H(T')$ at $H(t)$. In G, this translates to adding an edge, $\theta(H(E))$ of G that is connected to $\theta(H(T'))$ at $\theta(H(t))$, and then picking some $(k-1)$-subset of that new edge to be $\theta(H(s))$. When does this have property $\mathcal{P}$?

Consider $\theta'(T' \cup E)$. It is the same thing as $\theta'(T')$ but with an extra edge added, $\theta'(E)$, that is connected to it at $\theta'(t)$, and we also have a $(k-1)$-subset of that new edge, $\theta'(s)$. $\mathcal{P}$ requires that the image of $H' \circ \theta$ is a copy of $\theta'(T \cup E)$, and this implies that $\theta(H(s))$ matches $\theta'(s)$. There were $k-1$ choices for $\theta(H(s))$ and only one of them works. So out of all the $\deg(H(t))$ possible choices for the new edge, exactly $1/(k-1)$ of them have the right type of vertex at the other end.

However, that is not all we need to have property $\mathcal{P}$. It also needs to not self-intersect. To make sure that $\theta(H(T' \cup E))$ doesn't self-intersect, we need to make sure that when we picked $\theta(H(E))$, that it didn't intersect $H(T')$ anywhere except for $H(t)$. Since we are only adding 1 more vertex, we just need to make sure that this new vertex is not any of the vertices in $\theta(H(T'))\backslash\theta(H(t))$. That means there are at most $v(T')-k+1 \leq v(T)-1-k+1 = v(T)-k = e-1$ choices that make $\theta'(T_i)$ self-intersect.

Now if $\theta(H(s))$ matches $\theta'(s)$ and if $\theta(H(T' \cup E))$ doesn't self-intersect, then it is an actual copy of $\theta'(T' \cup E)$ and thus we satisfy property $\mathcal{P}$. All in all, this means there are at least $\deg(H(t_i))\frac{1}{k-1} - (e-1) = \deg(H(t_i))\frac{q_{T'\cup E}}{q_{T'}} - p$ possible choices for H'. So the last bullet point does indeed hold.

All this shows that $\mathcal{P}$ is weakly injective-like. But now given any homomorphism H, only $1/(k-1)$ of extensions of H have the right type of vertex at the other end. So there are at most $\deg(H(t))/(k-1) = \deg(H(t))\frac{q_{T'\cup E}}{q_{T'}}$ possible extensions of H. So $\mathcal{P}$ is in fact strongly injective-like. This means that we can apply theorem 2. It says that:

$$
\begin{aligned}
& |\operatorname{Hom}_{\mathcal{P}}(T, X)| \\
\geq\ & \binom{n}{k-1}(k-1)! \cdot d^e \cdot \frac{1}{(k-1)^e} \cdot \left(1 - e(e+2)(1+k-1)\ln\left(\binom{n}{k-1}(k-1)!\right)\frac{e-1}{d}\right) \\
\geq\ & \binom{n}{k-1}(k-1)! \cdot (d/(k-1))^e \cdot \left(1 - (e-1)e(e+2)(k-1)k\frac{\ln(n)}{d}\right)
\end{aligned}
$$

Chapter 3

Kruskal-Katona-type problem

3.1 Introduction

The Kruskal-Katona Theorem was proved in the 1960s by Kruskal and Katona [19, 17]. In the Theorem, we have families $\mathcal{A} \subset \mathbb{N}^{(r)}$ and $\mathcal{B} \subset \mathbb{N}^{(r-1)}$ such that for every $A \in \mathcal{A}$, there exist distinct $B_1, B_2, ..B_r \in \mathcal{B}$ with $B_i \subset A$. Given $a = |\mathcal{A}|$, we want to minimise $b = |\mathcal{B}|$, or equivalently, given b, we want to maximise a. The Kruskal Katona Theorem states that an optimal solution is when $\mathcal{A}$ is an initial segment of the colexicographic ordering on sets of size r, and $\mathcal{B}$ is the corresponding initial segment of the colexicographic ordering on sets of size $r-1$.

Definition 13 (Colexicographic ordering or colex)**.** *The colexicographic ordering of sets of size r is a total ordering where $A < B$ if the largest element of $B \backslash A$ is larger than the largest element of $A \backslash B$.*

Example: When $r = 4$, the first sets in the colex ordering are:
$\{1,2,3,4\}, \{1,2,3,5\}, \{1,2,4,5\}, \{1,3,4,5\}, \{2,3,4,5\}, \{1,2,3,6\}, \{1,2,4,6\}, \{1,3,4,6\}, ...$

More specifically, we can calculate what a or b is based on the other. To do so, we need to write them as a binomial sum. Pick some integers $c_{r-1} > c_{r-2} > c_{r-3} > ... > c_1$ in such a way as to make the following decomposition of b into binomials hold:

$$b = \binom{c_{r-1}}{r-1} + \binom{c_{r-2}}{r-2} + \binom{c_{r-3}}{r-3} + ... + \binom{c_1}{1}.$$

Then the maximum a is given by:

$$\binom{c_{r-1}}{r} + \binom{c_{r-2}}{r-1} + \binom{c_{r-3}}{r-2} + ... + \binom{c_1}{2}.$$

Conversely, given $a = \binom{c_{r-1}}{r} + \binom{c_{r-2}}{r-1} + \binom{c_{r-3}}{r-2} + ... + \binom{c_1}{2} + \binom{c_0}{1}$, the minimum for b is $\binom{c_{r-1}}{r-1} + \binom{c_{r-2}}{r-2} + \binom{c_{r-3}}{r-3} + ... + \binom{c_1}{1} + [1 \text{ if } c_0 > 0]$.

Note that, given any integers d and r, there exists a unique set of $c_r > c_{r-1} > c_{r-2} > ... > c_1$ that satisfy $d = \binom{c_{r-1}}{r-1} + \binom{c_{r-2}}{r-2} + \binom{c_{r-3}}{r-3} + ... + \binom{c_1}{1}$. Indeed, we have the formula $\binom{x}{s} + \binom{x-1}{s-1} + \binom{x-2}{s-2} + ... + \binom{x-s+1}{1} = \binom{x+1}{s} - 1$ for any x and s. So we can construct this set of c_is it by first picking c_{r-1} such that satisfies $\binom{c_{r-1}}{r-1} \leq d < \binom{c_{r-1}+1}{r-1}$. We subtract $\binom{c_{r-1}}{r-1}$ from d and then repeat the operation to find c_{r-2}. The remainder is between 0 and $\binom{c_{r-1}}{r-2} - 1$ so when we pick c_{r-2}, it will be strictly less than c_{r-1}. Repeat in the same way to find c_{r-3}, c_{r-4}, ... c_1. Therefore such a decomposition exists for all d.
As for uniqueness, we note that if we pick a different decomposition, there is some largest i where we picked different values for c_i. Say we use $c_i + t$ instead of c_i. Then $\binom{c_{r-1}}{r-1} + ... + \binom{c_i+t}{i} > d$ by definition of how we picked c_i to be the maximum number that worked. If on the other hand, we use $c_i - t$ instead of c_i, then the maximum number we can get is $\binom{c_{r-1}}{r-1} + ... + \binom{c_i-t+1}{i} - 1 < d$ so this also doesn't work. Therefore the choice of c_is is unique.

In 2015, Bollobás and Eccles [6] asked if the Kruskal-Katona Theorem could be generalised: instead of every subset of A being in $\mathcal{B}$, we instead required only k out of the r possible. In this case, we call the maximum value for a given r, k and b to be $f(r, k, b)$. They considered one configuration in particular: let $\mathcal{A}$ be of the form $\{S \cup X\}$ where X runs over an initial segment of the colex on sets of size k and S is just some set of size $r - k$ (that doesn't intersect any of the Xs). Meanwhile let $\mathcal{B}$ be defined in the same way as $\{S \cup Y\}$ where Y runs over the corresponding initial segment of the colex on sets of size $k - 1$. These collections of sets have the property that for every A in $\mathcal{A}$, there exist $B_1, B_2, ..., B_k$ in $\mathcal{B}$ with $B_i \subset A$.

This example gives a very similar formula to the Kruskal Katona Theorem. It shows that for any $c_{k-1} > c_{k-2} > ... > c_1$,

$$f\left(r, k, \binom{c_{k-1}}{k-1} + \binom{c_{r-2}}{k-2} + \binom{c_{k-3}}{k-3} + ... + \binom{c_1}{1}\right)$$
$$\geq \binom{c_{r-1}}{k} + \binom{c_{k-2}}{k-1} + \binom{c_{k-3}}{k-2} + ... + \binom{c_1}{2}.$$

Bollobàs and Eccles conjectured that this configuration is actually the optimal one when a and b are large. They did also note that this conjecture cannot be extended to small values of a and b, because they found an example that shows that $f(5,4,13) = 6$. If you tried to use the conjecture, it would tell you the answer is 7, which is incorrect. So the example is not optimal for small values of a and b; however, they still think this example is optimal when a and b are large enough.

In this chapter, the general methodology we will use will be to consider the set $\mathcal{B}$ to be the vertices of a k-hypergraph and $\mathcal{A}$ to be its edges. An vertex $B \in \mathcal{B}$ is incident to an edge $A \in \mathcal{A}$ is and only if $B \subset A$. We will first start by doing the easy cases of $k = 0$, $k = 1$, $k = 2$ and $k = 3$, which are all done using similar methods, although it gets more complicated as k increases.

Theorem 3. *(the cases where $k \leq 3$)*
- *For $0 = k \leq r$, then for all a, the minimum value for b is 0.*
- *For $1 = k \leq r$, the minimum value for b is 1 if $a \geq 1$, otherwise it is $b = 0$ if $a = 0$.*
- *For $2 = k \leq r$, $f(r,2,b) = \binom{b}{2}$.*
- *For $3 = k \leq r$, $f\left(r,3,\binom{c_2}{2} + \binom{c_1}{1}\right) = \binom{c_2}{3} + \binom{c_1}{2}$ whenever $c_2 > c_1$ and $c_2 \geq 29$.*

Note that this is still missing the cases where $c_2 < 29$; however there are only finitely many of these so they can in theory be solved by simply checking every single case. After this, we will move on to the case where $k \geq 4$. Unfortunately, there is a small gap in the proof which I only noticed late, so this case is not completely proved. The piece that is missing is that we need Conjecture 1 to be true in order to get the proof to work. It uses the same method, and if Conjecture 1 is true, it should end up giving exact results for an infinite number of values for b:

Theorem 4. *If Conjecture 1 is true (from chapter 2), then given $4 \leq k$, there is some constant μ depending only on k such that for all $r \geq k$, if $c_{k-1} > c_{k-2} > ... > c_1 > \mu$, then:*

$$f\left(r,k,\binom{c_{k-1}}{k-1} + ... + \binom{c_1}{1}\right) = \binom{c_{k-1}}{k} + ... + \binom{c_1}{2}.$$

Our method would also allow us to get to within some additive constant of the answer for all b:

Theorem 5. *If Conjecture 1 is true, then given $4 \leq k$, there is a constant τ depending*

only on k such that for all $r \geq k$, if $b = \binom{c_{k-1}}{k-1} + ... + \binom{c_1}{1}$, for some $c_{k-1} > c_{k-2} > ... > c_1$, then the maximum value for a is between:

$$\left[\binom{c_{k-1}}{k} + ... + \binom{c_1}{2}\right] \leq f(r,k,b) \leq \left[\binom{c_{k-1}}{k} + ... + \binom{c_1}{2}\right] + \tau.$$

Remark: Bollobás and Eccles also proposed the weaker conjecture that $f\left(r,k,\binom{x}{k-1}\right) \leq \binom{x}{k}$ whenever x is a positive real such that $\binom{x}{k-1}$ an integer. We do end up proving this in the cases $k \leq 3$. $k = 0, 1, 2$ are just corollaries of Theorem 3 while $k = 3$ is found during the proof of the Theorem 3. However, for the case $k \geq 4$, Theorems 4 and 5 are still the best we have so far.

3.2 The case $k = 0$

This one is trivial and you'll obviously have $\mathcal{B} = \emptyset$ regardless of what $\mathcal{A}$ is.

3.3 The case $k = 1$

This one is similarly trivial: the optimum will be $b = 1$ regardless of what $a \geq 1$ is. This can be achieved by letting $\mathcal{B}$ be an arbitrary set B of size $r - 1$, and $\mathcal{A}$ an arbitrary collection of a sets of size r all of which contain B.

3.4 The case $k = 2$

For any pair of elements in $\mathcal{B}$, there will be at most 1 element in $\mathcal{A}$ that contains both (which if it exists will be their union). Since every element of $\mathcal{A}$ does contain a pair of elements of $\mathcal{B}$, we have $|\mathcal{A}| \leq \binom{|\mathcal{B}|}{2}$ so $f(r,2,b) \leq \binom{b}{2}$.

This matches Bollobás's and Eccles's construction for the lower bound. Pick any some S of size $r - 2$ (all of whose elements are larger than b) and then define $\mathcal{B}$ to be $\{S \cup \{i\} : i \leq b\}$ and $\mathcal{A}$ to be $\{S \cup \{i,j\} : i,j \leq b\}$. This shows that $f(r,2,b) \geq \binom{b}{2}$ and therefore $f(r,2,b) = \binom{b}{2}$.

3.5 The case $k = 3$

Given a valid configuration $(\mathcal{A}, \mathcal{B})$, we can construct a k-hypergraph with b vertices corresponding to elements of $\mathcal{B}$. There are a edges corresponding to elements of $\mathcal{A}$; each one of these contains at least k elements of $\mathcal{B}$, and these k vertices are going to be the vertices the edge is incident to (if there are more than k of them, pick k of them arbitrarily).

Also, given two sets B_1 and B_2, we define the distance $d(B_1, B_2)$ between them as $|B_1 \triangle B_2|/2$ (so two adjacent vertices are at distance 1 from each other). Note that given 2 such sets at distance 1, there is at most one single edge that contains both: $B_1 \cup B_2$.

3.5.1 Number of paths of length 2

Consider our 3-hypergraph with b vertices and a edges. The average degree is $\frac{3a}{b}$.

Paths of length 2 Let the path of length 2, P_2 be the hypergraph consisting of 2 edges intersecting in a single point C (which we'll call the center), and with two distinguished points, one on each edge that are not the center: B_1 and B_2; we'll call these the endpoints.

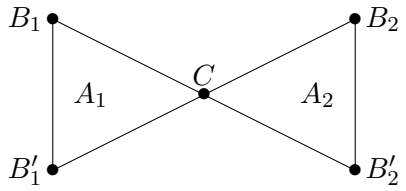


Figure 3.1: The hypergraph P_2

We want to count the number of P_2s. First, we'll pick out the center $C \in \mathcal{B}$. Then there are $deg(B)$ possible choices for A_1, then $deg(B) - 1$ choices for A_2. Finally, we have 2 choices each for B_1 and B_2. So overall, the number of P_2s is:

$$
\begin{aligned}
& |\{H : H \cong P_2\}| \\
= & \sum_{C \in \mathcal{B}} 4 \deg(C)(\deg(C) - 1) \\
= & \sum_{C \in \mathcal{B}} 4 \deg(C)^2 - \sum_{C \in \mathcal{B}} 4 \deg(C) \\
\geq & \frac{4 \left(\sum_{C \in \mathcal{B}} \deg(C) \right)^2}{\sum_{C \in \mathcal{B}} 1} - 4 \sum_{C \in \mathcal{B}} \deg(C) \qquad (3.1) \\
= & \frac{4(3a)^2}{b} - 4(3a) \\
= & \frac{36a^2}{b} - 12a.
\end{aligned}
$$

Now given any $C \in \mathcal{B}$, set $\epsilon_1(C) = 4 \left[\deg(C) - \sum_{B \in \mathcal{B}} \deg(B)/b \right]^2$. We have:

$$
\sum_{C \in \mathcal{B}} \epsilon_1(C) = 4 \sum_{C \in \mathcal{B}} \deg(C)^2 - 8 \left[\sum_{C \in \mathcal{B}} \deg(C) \right] \left[\frac{\sum_{B \in \mathcal{B}} \deg(B)}{b} \right] + 4b \left[\frac{\sum_{B \in \mathcal{B}} \deg(B)}{b} \right]^2
$$

$$
= 4 \sum_{C \in \mathcal{B}} \deg(C)^2 - 4 \frac{\left(\sum_C \deg(C) \right)^2}{b}.
$$

This is exactly the error in the above inequality (3.1), so the number of copies of P_2 is therefore exactly $\frac{36a^2}{b} - 12a + \sum_{C \in \mathcal{B}} \epsilon_1(C)$.

Paths of length 2 connecting points at distance 2 Now given a path of length 2, how often are the two endpoints at distance 2 and how often are they at distance 1? We claim that the number of copies of P_2 with endpoints at distance 2 is at least half of all copies of P_2. To see this, suppose we are given a copy of P_2 whose endpoints B_1 and B_2 (see Figure 1) are at distance 1. Then we can write $B_1 = C \cup \{x_1\} \backslash \{y_1\}$ and $B_2 = C \cup \{x_2\} \backslash \{y_1\}$. Then $B_2' = C \cup \{x_2\} \backslash \{y_2\}$ so B_2' and B_1 are at distance 2. Using a similar reasoning, B_2 and B_1' are at distance 2; however, B_1' and B_2' might still be at distance 1: we don't know. At any rate, for every copy of P_2 whose endpoints B_1 and B_2 are at distance 1, we can pick a unique corresponding copy of P_2 whose endpoints are B_1, B_1' at distance 2. Therefore at least half of all copies of P_2 have endpoints at distance 2, which is at least $\frac{18a^2}{b} - 6a + \sum_{C \in \mathcal{B}} \frac{\epsilon_1(C)}{2}$ copies.

Given a vertex C, we will define $\epsilon_2(C)$ to be the number of copies of P_2 having center C whose endpoints B_1 and B_2 are at distance 2 from each other but also that B_1, B_1' are at distance 2 from each other. With this addition we can be more precise and say that the number of copies of P_2 whose endpoints are at distance 2 is at least $\frac{18a^2}{b} - 6a + \sum_{C\in\mathcal{B}} \frac{\epsilon_1(C)}{2} + \sum_{C\in\mathcal{B}} \frac{\epsilon_2(C)}{2}$.

Distance 2 pairs Now given two points B and B' at distance 2, there are at most 4 paths of length 2 from B to B' (because there are at most 4 possible centres). Therefore the number of ordered pairs of points (B, B') is at least the number of paths from one vertex to another at distance 2 from it, divided by 4.

Given $B \in \mathcal{B}$, let $\epsilon_3(B)$ to be the number of points B' at distance 2 from B such that there are 3 or fewer paths of length 2 from B to B'. Using this, we get:

$$|\{H : H \cong P_2; d(B_1, B_2) = 2\}| \geq 4|\{(B, B') : d(B, B') = 2\}| - 4\sum_{B\in\mathcal{B}} \epsilon_3(B).$$

And thus the total number of ordered pairs of points (B, B') at distance 2 from each other is:

$$\begin{aligned} & |\{(B, B') : d(B, B') = 2\}| \\ \geq \quad & \frac{9a^2}{2b} - \frac{3a}{2} + \sum_{C\in\mathcal{B}} \frac{\epsilon_1(C)}{8} + \sum_{C\in\mathcal{B}} \frac{\epsilon_2(C)}{8} + \sum_{C\in\mathcal{B}} \epsilon_3(C). \end{aligned}$$

We also know that the number of ordered pairs of points (B, B') at distance 1 is at least $6a$ (6 from each edge). We set $\epsilon_4(B)$ to be the number of points at distance 1 from B such that there is no edge connecting them. So $6a + \sum_B \epsilon_4(B)$ is the total number of ordered pairs at distance 1. Finally, let $\epsilon_5(B)$ be the number of points at distance at least 3 from B, so $\sum_B \epsilon_5(B)$ is the total number of ordered pairs of points at distance at least 3. Together with our calculated 'number of ordered pairs at distance 2', this accounts for every possible ordered pair of points. We know that the total number of such pairs is $b(b-1)$. Thus:

$$b(b-1) \geq \frac{9a^2}{2b} - \frac{3a}{2} + 6a + \sum_{C \in \mathcal{B}} \left[\frac{\epsilon_1(C)}{8} + \frac{\epsilon_2(C)}{8} + \epsilon_3(C) + \epsilon_4(C) + \epsilon_5(C) \right].$$

We know that all the $\epsilon_i(B)$s are non-negative so we get the inequality:

$$b(b-1) \geq \frac{9a^2/b - 3a}{2} + 6a = \frac{9a^2/b + 9a}{2}.$$

Solving this for a gives:

$$a \leq \frac{-3 + \sqrt{8b+1}}{6} b. \tag{3.1}$$

When we set $b = \binom{c}{2}$, this inequality gives us $a \leq \binom{c}{3}$. This is tight. Indeed, we can look at the configuration from the hypothesis: let S be a set of size $r-3$ (that doesn't contain any of 1,2,3,...,c) and let $\mathcal{B} = \{S \cup \{i,j\} | i,j \leq c\}$ and $\mathcal{A} = \{S \cup \{i,j,k\} | i,j,k \leq c\}$. Then this is a valid configuration and has $b = \binom{c}{2}$ and $a = \binom{c}{3}$.

Remark: This formula is exactly the same as $f\left(r, 3, \binom{x}{2}\right) \leq \binom{x}{3}$ for every real $x \geq 3$ such that $\binom{x}{2}$ is a positive integer, thereby proving the weaker conjecture in the case $k = 3$.

3.5.2 The existence of a large nice hypergraph when a is close to the upper bound

For this section, we will assume that $c_2 \geq 29$. If we set $b = \binom{c_2}{2} + c_1$ and $a = \binom{c_2}{3} + \binom{c_1}{2} + 1$ for some $c_1 < c_2$, then this means that the sum of all the $\epsilon_i(C)$ are small. In fact, we get:

$$\frac{\sum_{C \in \mathcal{B}}}{b} \left[\frac{\epsilon_1(C)}{8} + \frac{\epsilon_2(C)}{8} + \epsilon_3(C) + \epsilon_4(C) + \epsilon_5(C) \right]$$
$$= (b-1) - \frac{9a^2}{2b^2} - \frac{9a}{2b}.$$

This is an average, so in particular, there exists a vertex B for which :

$$\frac{\epsilon_1(B)}{8} + \frac{\epsilon_2(B)}{8} + \epsilon_3(B) + \epsilon_4(B) + \epsilon_5(B) \leq (b-1) - \frac{9a^2}{2b^2} - \frac{9a}{2b}.$$

And for brevity, we'll set

$$\gamma = (b-1) - \frac{9a^2}{2b^2} - \frac{9a}{2b} = \frac{c_2^2 - c_2 + 2c_1 - 2}{2} - \frac{9}{2}\left(\frac{a}{b}\right)^2 - \frac{9}{2}\left(\frac{a}{b}\right) \tag{3.2}$$

.

We will try to bound γ from above so as to guarantee having small ϵs

We have $\frac{a}{b} = \frac{1}{3}\frac{c_2^3 - 3c_2^2 + 2c_2 + 3c_1^2 - 3c_1 + 6}{c_2^2 - c_2 + 2c_1} = \frac{c_2 - 2 - 2c_1/c_2 + 3c_1^2/c_2^2}{3} + \frac{1}{3}\frac{-c_1 + 7c_1^2/c_2 - 6c_1^3/c_2^2 + 6}{c_2^2 - c_2 + 2c_1}$. But $-c_1 + 7c_1^2/c_2 - 6c_1^3/c_2^2 + 6$ can be written as $6 + c_1(-1 + 7z - 6z^2)$ for $z = c_1/c_2$ which is between 0 and 1. The minimum of the function $-1 + 7z - 6z^2$ for z between 0 and 1 is -1, and therefore $\frac{a}{b} \geq \frac{c_2 - 2 - 2c_1/c_2 + 3c_1^2/c_2^2}{3} + \frac{1}{3}\frac{6 - c_1}{c_2^2 - c_2 + 2c_1}$. And now as long as $c_1 \geq 7$, we have $0 > \frac{6 - c_1}{c_2^2 - c_2 + 2c_1} \geq \frac{6 - (c_2 - 1)}{c_2^2 - c_2} = \frac{7 - c_2}{c_2(c_2 - 1)} \geq -\frac{1}{c_2}$. So we know that $\frac{a}{b} \geq \frac{c_2 - 2 - 2c_1/c_2 + 3c_1^2/c_2^2}{3} - \frac{1}{3c_2}$. If $c_1 \leq 6$, then $\frac{6 - c_1}{c_2^2 - c_2 + 2c_1}$ is positive or zero so $\frac{a}{b} \geq \frac{c_2 - 2 - 2c_1/c_2 + 3c_1^2/c_2^2}{3}$. Regardless of which case we're in, we will have $\frac{a}{b} \geq \frac{c_2 - 2 - 2c_1/c_2 + 3c_1^2/c_2^2 - 1/7}{3}$

Using this, we can also say that if $c_1 \geq 7$, that

$$\begin{aligned}
&\left(\frac{a}{b}\right)^2 \\
&\left(\frac{c_2 - 2 - 2c_1/c_2 + 3c_1^2/c_2^2}{3} - \frac{1}{3c_2}\right)^2 \\
= \;& \frac{c_2^2 - 4c_2 - 4c_1 + 6c_1^2/c_2 + 4 + 8c_1/c_2 - 8c_1^2/c_2^2 - 12c_1^3/c_2^3 + 9c_1^4/c_2^4}{9} \\
& - \frac{2 - 4/c_2 - 4c_1/c_2^2 + 6c_1^2/c_2^3}{9} + \frac{1}{9c_2^2} \\
= \;& \frac{c_2^2 - 4c_2 - 4c_1 + 6c_1^2/c_2}{9} + \frac{2 + 8c_1/c_2 - 8c_1^2/c_2^2 - 12c_1^3/c_2^3 + 9c_1^4/c_2^4}{9} \\
& + \frac{4 + 4c_1/c_2 - 6c_1^2/c_2^2}{9c_2} + \frac{1/c_2^2}{9}
\end{aligned}$$

Notice that the function $2 + 8z - 8z^2 - 12z^3 + 9z^4 = (3z^2 - 2z - 2)^2 - 2 \geq -2$; taking $z = c_1/c_2$ gives a lower bound of $\frac{-2}{9}$ for the second term of the above inequality. We also look at the function $4 + 4z - 6z^2$ which is concave and valued at 4 when $z = 0$ and 2 when $z = 1$; so this function is always at least 2 when $0 \leq z \leq 1$; $0 \leq c_1/c_2 \leq 1$ so setting $z = c_1/c_2$ gives a lower bound of $\frac{2}{9c_2}$ for the third term of the above inequality. Using these, we end up with:

$$\left(\frac{a}{b}\right)^2 \geq \frac{c_2^2 - 4c_2 - 4c_1 + 6c_1^2/c_2}{9} - \frac{2}{9} + \frac{2}{9c_2} + \frac{1}{9c_2^2}.$$

And finally:

$$\left(\frac{a}{b}\right)^2 \geq \frac{c_2^2 - 4c_2 - 4c_1 + 6c_1^2/c_2 - 2}{9}.$$

In the case where $c_1 \leq 6$, we get:

$$\begin{aligned}
& \left(\frac{a}{b}\right)^2 \\
\geq\ & \frac{c_2^2 - 4c_2 - 4c_1 + 6c_1^2/c_2 + 4 + 8c_1/c_2 - 8c_1^2/c_2^2 - 12c_1^3/c_2^3 + 9c_1^4/c_2^4}{9} \\
=\ & \frac{c_2^2 - 4c_2 - 4c_1 + 6c_1^2/c_2}{9} + \frac{1}{9}\left(3\frac{c_1^2}{c_2^2} - 2\frac{c_1}{c_2} - 2\right)^2 \\
\geq\ & \frac{c_2^2 - 4c_2 - 4c_1 + 6c_1^2/c_2}{9}
\end{aligned}$$

So regardless of whether $c_1 \geq 7$ or $c_1 \leq 6$, we know that

$$\left(\frac{a}{b}\right)^2 \geq \frac{c_2^2 - 4c_2 - 4c_1 + 6c_1^2/c_2 - 2}{9}.$$

Now we go back to γ, which was, if you recall from 3.2, equal to $\frac{c_2^2 - c_2 + 2c_1 - 2}{2} - \frac{9}{2}\left(\frac{a}{b}\right)^2 - \frac{9}{2}\left(\frac{a}{b}\right)$. Now that we have good bounds on $\frac{a}{b}$ and $\left(\frac{a}{b}\right)^2$, we can find a good bound on γ:

$$\begin{aligned}
\gamma =\ & \frac{c_2^2 - c_2 + 2c_1 - 2}{2} - \frac{9}{2}\left(\frac{a}{b}\right)^2 - \frac{9}{2}\left(\frac{a}{b}\right) \\
\leq\ & \frac{c_2^2 - c_2 + 2c_1 - 2}{2} - \frac{c_2^2 - 4c_2 - 4c_1 + 6c_1^2/c_2 - 2}{2} - \frac{3c_2 - 6 - 6c_1/c_2 + 9c_1^2/c_2^2 - 3/7}{2} \\
=\ & \frac{6c_1 - 6c_1^2/c_2 + 45/7 + 6c_1/c_2 - 9c_1^2/c_2^2}{2} \\
=\ & \frac{6c_1 - 6c_1^2/c_2 + 52/7}{2} - \frac{(3c_1/c_2 - 1)^2}{2} \\
<\ & 3c_1 - 3c_1^2/c_2 + 4 \\
\leq\ & \frac{3c_2}{4} + 4 \\
\leq\ & c_2
\end{aligned}$$

So as long as $c_2 \geq 16$, we know that $\gamma < c_2$. We assumed at the start that $c_2 \geq 29$

so we will indeed have $\gamma < c_2$.

Using γ to restrict what the families looks like We know from the definition of γ (3.2) that $\epsilon_1(B) \leq 8\gamma$ so using the definition of ϵ_1, we get $4|\deg(B) - 3a/b|^2 \leq 8c_2$ so $|\deg(B) - 3a/b| \leq \sqrt{2c_2}$. The degree of B is therefore pretty close to its expected value.

Now we also know that the number of points B' at distance 2 from B is $b - 1 - 2\deg(B) - \epsilon_4(B) - \epsilon_5(B)$ since $\epsilon_4(B)$ is the number of points at distance 1 from B and $\epsilon_5(B)$ is the number of points at distance at least 3. We want to count the number of such points that have all 4 possible paths between it and B. This is just $b - 1 - 2\deg(B) - \epsilon_4(B) - \epsilon_5(B) - \epsilon_3(B)$.

Again using the definition of γ (3.2), we know that $\epsilon_3(B) + \epsilon_4(B) + \epsilon_5(B)$ is less than γ and thus less than c_2 so the number of points B' at distance 2 from B with all 4 possible paths between B and B' is at least $b - 1 - 2\deg(B) - c_2$. We'll call such a configuration an <u>octahedron</u>.

We'll say that the $\deg(B)$ edges adjacent to B are $B \cup \{x_1\}, B \cup \{x_2\}, ...B \cup \{x_{\deg(B)}\}$, and say that the $2\deg(B)$ vertices at distance 1 from B on these edges are: $B \cup \{x_1\} \backslash \{y_1\}, B \cup \{x_1\} \backslash \{z_1\}, B \cup \{x_2\} \backslash \{y_2\}, B \cup \{x_1\} \backslash \{z_2\}, ..., B \cup \{x_{\deg(B)}\} \backslash \{y_{\deg(B)}\}, B \cup \{x_{\deg(B)}\} \backslash \{z_{\deg(B)}\}$

We'll now colour the edges incident to B as follows: every edge incident to B has vertices of the form B,$B \cup \{x\} \backslash \{y\}$ and $B \cup \{x\} \backslash \{z\}$. The colour of this edge is defined to be $\{y, z\}$. This colouring is useful because of its relationship to the octahedrons.

Indeed, an octahedron is formed of the 6 points: $B, B \cup \{x_i\} \backslash \{y_i\}, B \cup \{x_i\} \backslash \{z_i\}, B \cup \{x_j\} \backslash \{y_j\}, B \cup \{x_j\} \backslash \{z_j\}$ and $B \cup \{x_i, x_j\} \backslash \{y_i, z_i\}$ and also requires that $\{y_i, z_i\} = \{y_j, z_j\}$. That means that each octahedron contains 2 edges of the same colour. Furthermore, given a pair of edges of the same colour, there can only be at most 1 octahedron that contains both. So the number of octahedrons is smaller than the number of pairs of edges of the same colour.

We'll define s to be the size of the largest colour class. How large must s be?

If s is fixed and $s \geq \deg(B)/2$, then the maximum number of octahedrons is $\binom{s}{2} + \binom{\deg(B)-s}{2} = s^2 + \frac{\deg(B)^2 - \deg(B)}{2} - \deg(B)s$. We know that this quantity is at least as large

as the actual number of octahedrons, which is in turn larger than $b - 1 - 2\deg(B) - c_2$. So we have:

$$s^2 - \deg(B)s + \frac{-2b + 2 + 2c_2 + 3\deg(B) + \deg(B)^2}{2} \geq 0.$$

We will use this inequality to find a lower bound on s. First, however, we need to find what $\deg(B)$ can be. Note that if we differentiate $s^2 - \deg(B)s + \frac{-2b+2+2c_2+3\deg(B)+\deg(B)^2}{2} \geq 0$ by $\deg(B)$, we get $-s + \frac{3}{2} + \deg(B)$. Since $s \leq \deg(B)$, this is always positive and therefore we can without loss of generality assume that we are in the worst case scenario where $\deg(B)$ is the maximum it can possibly be.

We know one upper bound on $\deg(B)$: $\deg(B) \leq 3a/b + \sqrt{2c_2}$, but it isn't very nice to work with. So we simplify $\frac{3a}{b} = \frac{c_2^3 - 3c_2^2 + 2c_2 + 3c_1^2 - 3c_1}{c_2^2 - c_2 + 2c_1} = c_2 - 2 - 2c_1/c_2 + 3c_1^2/c_2^2 + \frac{-c_1 + 7c_1^2/c_2 - 6c_1^3/c_2^2 + 6}{c_2^2 - c_2 + 2c_1} = c_2 - 2 + (-2c_1/c_2 + 3c_1^2/c_2^2) + \frac{6 + c_2(-c_1/c_2 + 7c_1^2/c_2^2 - 6c_1^3/c_2^3)}{c_2^2 - c_2 + 2c_1}$. Looking at the function $2z + 3z^2$ for $0 \leq z \leq 1$,this is has a maximum of 1 at $z = 1$. Meanwhile, the function $-z + 7z^2 - 6z^3$ is smaller than 1 on the same interval. Therefore $\frac{3a}{b} \leq c_2 - 2 + 1 + \frac{6 + c_2}{c_2^2 - c_2 + 2c_1} \leq c_2 - 1 + \frac{6+29}{29^2 - 29} \leq c_2$.

So we have an upper bound for $\deg(B)$: $\deg(B) \leq c_2 + \sqrt{2c_2}$. And remember that we could assume without loss of generality that $\deg(B)$ was maximal, so we set $\deg(B) = c_2 + \sqrt{2c_2}$. We were trying to find a lower bound on s using the inequality $s^2 - \deg(B)s + \frac{-2b+2+2c_2+3\deg(B)+\deg(B)^2}{2} \geq 0$. This is a quadratic which we can solve:

$$\begin{aligned}
s \geq & \frac{\deg(B) + \sqrt{4b - 4 - 4c_2 - 6\deg(B) - \deg(B)^2}}{2} \\
= & \frac{c_2 + \sqrt{2c_2} + \sqrt{2c_2^2 - 2c_2 + 4c_1 - 4 - 4c_2 - 6(c_2 + \sqrt{2c_2}) - (c_2 + \sqrt{2c_2})^2}}{2} \\
= & \frac{c_2 + \sqrt{2c_2} + \sqrt{2c_2^2 - 12c_2 + 4c_1 - 4 - 6\sqrt{2c_2} - c_2^2 - (2c_2)^{3/2} - 2c_2}}{2} \\
= & \frac{c_2 + \sqrt{2c_2} + \sqrt{c_2^2 - 14c_2 + 4c_1 - 4 - 6\sqrt{2c_2} - (2c_2)^{3/2}}}{2} \\
\geq & \frac{c_2 + \sqrt{2c_2} + \sqrt{c_2^2 - 14c_2 - 4 - 6\sqrt{2c_2} - (2c_2)^{3/2}}}{2}
\end{aligned}$$

$$
\begin{aligned}
&= \frac{c_2 + \sqrt{2c_2} + \sqrt{(c_2 - \sqrt{2c_2} - 8)^2 - 22\sqrt{2c_2} - 68}}{2} \\
&\geq \frac{c_2 + \sqrt{2c_2} + \sqrt{(c_2 - \sqrt{2c_2} - 8)^2}}{2} \\
&= \frac{2c_2 - 8}{2} \\
&= c_2 - 4
\end{aligned}
$$

So this colour class of size s encompasses most of the neighbourhood of B when $c_2 \geq 29$. In fact, there are at most $\deg(B) + 4 - c_2$ points of the neighbourhood that are outside. Since $\deg(B) \leq c_2 + \sqrt{2c_2}$, that means there are at most $4 + \sqrt{2c_2}$ points in the neighbourhood of B outside our colour class. We'll say that the colour of this large colour class is $\{y, z\}$. Finally, we'll define $S = B \backslash \{y, z\}$ and define the *nice hypergraph* to be the set of all vertices and edges that contain S as a subset. Our nice hypergraph contains our large colour class as well as all the octahedrons related to it.

We ask ourselves how many octahedrons do we have in the nice hypergraph? We know that there are at least $b - 1 - 2\deg(B) - c_2$ octahedrons that contain B in total. The maximum number of octahedrons containing B that are not in the nice hypergraph is $\binom{\deg(B)-s}{2} \leq \binom{4+\sqrt{2c_2}}{2} = \frac{12+7\sqrt{2c_2}+2c_2}{2}$. Therefore the number of octahedrons in our big colour class is at least $b-1-2\deg(B)-c_2-\frac{12+7\sqrt{2c_2}+2c_2}{2} = b-7-2\deg(B)-2c_2-7\sqrt{c_2/2}$. That's almost all the vertices in the graph! Remembering the $2\deg(B)$ vertices adjacent to B and B itself, that means there are only $6 + 7\sqrt{c_2/2} + 2c_2$ vertices left unaccounted for. Since $c_2 \geq 29$, this is less than $4c_2$. Before we finish up the proof, we will need one more lemma, which also covers the example case where $b = \binom{c_2}{2}$ and $a = \binom{c_2}{3}$.

3.5.3 Example case where $b = \binom{c_2}{2}$ and $a = \binom{c_2}{3}$

Lemma 2. *The optimal configuration when $k = 3$, $b = \binom{c_2}{2}$ and $a = \binom{c}{3}$ is of the form:*
$\mathcal{A} = \{S \cup T | T \in \{x_1, x_2, ..., x_{c_2}\}^{(3)}\}$
$\mathcal{B} = \{S \cup T | T \in \{x_1, x_2, ..., x_{c_2}\}^{(2)}\}$.

Proof:

First of all, since there was equality in this case, that means that $\gamma = 0$ and thus all the error terms ϵ_i are also zero. This implies that the degree of every vertex is exactly $3a/b = c_2 - 2$ and for every vertex B, the number of vertices at distance 2 from it

is exactly $b - 1 - 2\deg(B)$. Finally, we know that the every single one of these points forms an octahedron with B and 2 points in the neighbourhood of B.

We can write the neighbourhood of B as $B \cup \{x_i\}\backslash\{y\}$ and $B \cup \{x_i\}\backslash\{z\}$ for all i between 1 and $c_2 - 2$. Finally, we can write every other point of the graph in the form $B \cup \{x_i, x_j\}\backslash\{y, z\}$ for all $1 \leq i < j \leq c_2 - 2$.

Let $S = B \cup \backslash\{y\}$ and then $\mathcal{A}$ is exactly the family $\{S \cup \{t_1, t_2, t_3\} | t_1, t_2, t_3 \in \{y, z, x_1, x_2, ..., x_{c_2-2}\}\}$ while $\mathcal{B}$ is exactly the family $\{S \cup \{t_1, t_2\} | t_1, t_2 \in \{y, z, x_1, x_2, ..., x_{c_2-2}\}\}$.

□

3.5.4 Other cases

As a reminder, at this stage we know that there is a large 'nice hypergraph' of vertices that all contain S as a subset. We'll say that there are exactly β vertices not in our nice hypergraph (which leaves $b - \beta$ vertices in the nice hypergraph). We know that $\beta < 4c_2$. How many edges can there be in this graph now that we have this information? There are 3 types of edges, depending on how many vertices are in the nice hypergraph:

• The edges that are entirely contained in the nice hypergraph. We shall call these nice edges. All the vertices in the nice hypergraph contain S so by the Kruskal Katona Theorem, the most edges entirely contained within is when they form an initial segment of the colex ordering. So if we write $b - \beta = \binom{d_2}{2} + \binom{d_1}{1}$, then we have at most $\binom{d_2}{3} + \binom{d_1}{2}$ nice edges. Because $\beta < 4c_2$ and $c_2 \geq 29$, we have $b - \beta > \frac{c_2^2 - c_2 + 2c_1}{2} - 4c_2 \geq \frac{c_2^2 - 9c_2}{2} \geq \binom{c_2 - 5}{2}$. Therefore $d_2 \geq c_2 - 5$.
• The edges that that contain 1 or 2 vertices from the nice hypergraph and 2 or 1 from outside. We shall call these linking edges. There are at most β of them because given any point T outside the nice hypergraph, the only potential edge that can connect to elements in the nice hypergraph is $T \cup S$.
• The edges entirely outside the nice hypergraph. We'll call these outside edges. If we just apply our earlier result (3.1), we get that there are at most $\frac{-3+\sqrt{8\beta+1}}{6}\beta$ of them.

In total, we have at most $\binom{d_2}{3} + \binom{d_1}{2} + \frac{3+\sqrt{8\beta+1}}{6}\beta$ edges in our hypergraph. We shall try to go for a contradiction and assume the number of edges is also equal to $\binom{c_2}{3} + \binom{c_1}{2} + 1$. This implies that:

$$(c_2 - d_2)\frac{c_2^2 + c_2 d_2 + d_2^2 - 3c_2 - 3d_2 + 2}{6} + (c_1 - d_1)\frac{c_1 + d_1 - 1}{2} \leq \frac{3 + \sqrt{8\beta + 1}}{6}\beta.$$

And $\beta = (c_2 - d_2)\frac{c_2 + d_2 - 1}{2} + (c_1 - d_1)$. Let $\phi = c_2 - d_2$; we know that $0 \leq \phi \leq 4$. Replace all instances of d_2 with $c_2 - \phi$ in the inequality. We end up with:

$$\phi\left(3c_2^2 - 3\phi c_2 + \phi^2 - 6c_2 + 3\phi + 2\right) + 3(c_1 - d_1)(c_1 + d_1 - 1) \leq (3 + \sqrt{8\beta + 1})\beta. \quad (3.3)$$

And $\beta = \phi\left(c_2 - \frac{\phi+1}{2}\right) + (c_1 - d_1)$.

Case 1: $2 \leq \phi \leq 5$

We know that $0 \leq c_1 \leq c_2 - 1$ and $0 \leq d_1 \leq d_2 - 1$ so $(c_1 - d_1)(c_1 + d_1 - 1) \geq -d_1(d_1 - 1) \geq -(d_2 - 1)(d_2 - 2) = -(c_2 - \phi - 1)(c_2 - \phi - 2) = -(c_2 - 1)(c_2 - 2) + \phi(2c_2 - 3) - \phi^2$ and $\beta \leq \phi\left(c_2 - \frac{\phi+1}{2}\right) + c_2 - 1 = (\phi + 1)(c_2 - \phi/2) - 1$. Putting these back into inequality 3.3, we get:

$$\begin{aligned} & \phi\left(3c_2^2 - 3\phi c_2 + \phi^2 - 6c_2 + 3\phi + 2\right) - 3(c_2 - 1)(c_2 - 2) + 3\phi(2c_2 - 3) - 3\phi^2 \\ \leq\ & (3 + \sqrt{8(\phi + 1)(c_2 - \phi/2) - 7})[(\phi + 1)(c_2 - \phi/2) - 1]. \end{aligned}$$

So:

$$\begin{aligned} & (3\phi - 3)c_2^2 + (-3\phi^2 + 9)c_2 + (\phi^3 - 7\phi - 6) \\ \leq\ & 3(\phi + 1)(c_2 - \phi/2) - 3 + \sqrt{8(\phi + 1)(c_2 - \phi/2) - 7}[(\phi + 1)(c_2 - \phi/2) - 1]. \end{aligned}$$

Therefore:

$$\begin{aligned} & (3\phi - 3)c_2^2 + (-3\phi^2 - 3\phi + 6)c_2 + (\phi^3 + 3/2\phi^2 - 11/2\phi - 3) \\ \leq\ & \sqrt{8(\phi + 1)(c_2 - \phi/2) - 7}[(\phi + 1)(c_2 - \phi/2) - 1]. \end{aligned}$$

For $\phi = 2$, this gives us $3c_2^2 - 12c_2 \leq \sqrt{24c_2 - 31}(3c_2 - 4)$ which is false for $c_2 \geq 29$.

For $\phi = 3$, this gives us $6c_2^2 - 30c_2 + 21 \leq \sqrt{32c_2 - 55}(4c_2 - 7)$ which is false for $c_2 \geq 20$.

For $\phi = 4$, this gives us $9c_2^2 - 54c_2 + 63 \leq \sqrt{40c_2 - 87}(5c_2 - 11)$ which is false for $c_2 \geq 18$

For $\phi = 5$, this gives us $12c_2^2 - 84c_2 + 132 \leq \sqrt{48c_2 - 127}(6c_2 - 16)$ which is false for $c_2 \geq 18$

We assumed that $c_2 \geq 29$ so therefore none of these cases can occur.

Case 2: $\phi = 1$

Then inequality 3.3 becomes:

$$\left(3c_2^2 - 9c_2 + 6\right) + (c_1 - d_1)\left(3c_1 + 3d_1 - 3\right) \leq (3 + \sqrt{8\beta + 1})\beta. \tag{3.4}$$

where $\beta = c_2 - 1 + (c_1 - d_1) \leq 2c_2 - 2$ so that implies:

$$\left(3c_2^2 - 12c_2 + 9\right) + 3c_1^2 - 3d_1^2 + 6(d_1 - c_1) \leq 8c_2^{3/2}.$$

But now $3c_2^2 - 8c_2^{3/2} - 12c_2 + 9 \geq 3(c_2 - 4/3\sqrt{c_2} - 4)^2$ so this implies that $d_1 \geq c_2 - 4/3\sqrt{c_2} - 4$. We also get that $c_1 \leq \frac{1}{3}\sqrt{8c_2^{3/2} + 12c_2 - 9} < c_2/2$.

Replacing $d_1 = c_2 - \epsilon$ in inequality 3.4, we get:

$$(6\epsilon - 6)c_2 + 3c_1^2 - 6c_1 - 3\epsilon^2 - 6\epsilon + 9 \leq \sqrt{8c_1 + 8\epsilon - 7}(c_1 + \epsilon - 1).$$

And we know that $2 \leq \epsilon \leq 4/3\sqrt{c_2} + 4 < c_2/2$. Since $\sqrt{8c_1 + 8\epsilon - 7} < \sqrt{8c_2}$, we can say that:

$$-3\epsilon^2 + \epsilon(6c_2 - \sqrt{8c_2} - 6) + 3(c_1 - 1)^2 - \sqrt{8c_2}(c_1 - 1) - 6c_2 + 6 \leq 0.$$

The value for c_1 that minimises the left hand side is $c_1 = 1 + \sqrt{2c_2}/3$ so we can without loss of generality assume that is what c_1 is, and that gives us:

$$-3\epsilon^2 + \epsilon(6c_2 - \sqrt{8c_2} - 6) - 20/3c_2 + 6 \leq 0.$$

If we set ϵ to be at its maximum value: $\epsilon = 4/3\sqrt{c_2} + 4$, we get the left hand side is $8c_2^{3/2} + (12 - 8\sqrt{2}/3)c_2 - (40 + 8\sqrt{2})\sqrt{c_2} - 2$ which is positive for $c_2 \geq 29$. Similarly, when

ϵ is at its minimum value, $\epsilon = 2$, we have $-18 + 16/3c_2 - 4\sqrt{2c_2}$ which is also positive for $c_2 \geq 29$. As this function is a quadratic polynomial with a negative leading term, it is concave and therefore this inequality does not hold for any valid ϵ. So therefore this case cannot occur.

Case 3: $c_2 = d_2$ **but** $c_1 \neq d_1$

Then we can simplify (3.3) further to $(3c_1 + 3d_1 - 6) \leq \sqrt{8c_1 - 8d_1 + 1}$ so either $c_1 + d_1 \leq 2$ or $3c_1 - 6 \leq \sqrt{8c_1 + 1}$ so $9c_1^2 - 43c_1 + 35 \leq 0$ so $c_1 \leq 3$ in either case. Looking at each subcase individually, we get the following cases:

• Case 3.1. $d_1 = 0$ and $c_1 = 3$. In this case, our potential counter-example consists of 3 vertices outside the nice hypergraph with one outside edge and three linking edges (one per vertex), which is one more than the $\binom{3}{2} = 3$ we expected. But this configuration is actually impossible.
Indeed, if we remove these 3 outside vertices and their 4 edges, we're left with a $\binom{c_2}{2}$ vertices and $\binom{c_2}{3}$ edges so by Lemma 2, there is only a single unique solution: $\mathcal{B} = \{S \cup T | T \in \{x_1, x_2, ..., x_{c_2}\}^{(2)}\}$ and $\mathcal{A} = \{S \cup T | T \in \{x_1, x_2, ..., x_{c_2}\}^{(3)}\}$.
Also note that each of the 3 linking edges is only incident to one of the outside vertices, which means that its two other endpoints are inside the nice hypergraph. Say they are $S \cup \{x_1, x_2\}$ and $S \cup \{x_1, x_3\}$. Then the linking edge has to be $S \cup \{x_1, x_2, x_3\}$. But this is one of the nice edges that we've already counted. Contradiction. Therefore this configuration is indeed impossible.

• Case 3.2. $d_1 = 0$ and $c_1 = 2$. In this case, our potential counter example consists of 2 vertices outside the nice hypergraph with two linking edges, which is one more than the $\binom{2}{2} = 1$ we expected. This configuration is also impossible.
Similarly to the last bullet point, we note that if we remove the 2 extra vertices and edges, we are left with families of the type: $\mathcal{B} = \{S \cup T | T \in \{x_1, x_2, ..., x_{c_2}\}^{(2)}\}$ and $\mathcal{A} = \{S \cup T | T \in \{x_1, x_2, ..., x_{c_2}\}^{(3)}\}$. But also each of the 2 linking edges has to be of the form $S \cup \{x_1, x_2, x_3\}$, which is not a linking edge at all, but rather a nice edge that we have already counted. Contradiction.

• Case 3.3. $d_1 = 0$ and $c_1 = 1$. In this case, our potential counter example consists of 1 vertex outside the nice hypergraph with one linking edges, which is one more than the $\binom{1}{2} = 0$ we expected. This configuration is also impossible and the proof is identical to the last two bullet points.

• Case 3.4. $d_1 = 1$ and $c_1 = 2$. In this case, our potential counterexample can actually work. However it has $\binom{c_2}{3} + \binom{1}{2} + 1$ edges which is the same as $\binom{c_2}{3} + \binom{2}{2}$ that our usual example gives us so it doesn't give any improvement.

So in conclusion, we have proved the conjecture in the case $k = 3$ and for $c_2 \geq 29$. This finishes the proof of Theorem 3.

Remark: we did not answer the question of what happens when $c_2 < 29$; however, this is only finitely many cases so it could in theory be solved by simply checking all the cases individually.

3.6 The case $k \geq 4$

In this section, we are going to prove Theorems 4 and 5. Unfortunately, it does require Conjecture 1 (from Chapter 2) to be true instead of Theorem 2 as was expected. Similarly to the case $k = 3$, we again define the distance $d(B_1, B_2) = |B_1 \triangle B_2|/2$.

We will follow what we did in the case $k = 3$ except we will be looking at pairs of points at distance $k - 1$ from each other, instead of pairs at distance 2, and the paths joining them together will have length $k - 1$ instead of length 2.

Here, we will assume that conjecture 1 is true and use it. The graph X will be the one that has $\mathcal{B}$ as vertices. We know that every element A of $\mathcal{A}$ contains at least k elements of $\mathcal{B}$ as subsets, so pick k of them arbitrarily, then connect these k vertices by edges. Thus, the graph has exactly $ak(k-1)/2$ edges, making the average degree of this graph be $k(k-1)\frac{a}{b}$.

The tree T will be a path P_{k-1} of length $k - 1$. The sequence of subtrees we use to construct it will just be made up of all the shorter paths: $P_0 \subset P_1 \subset P_2 \subset ... \subset P_{k-1}$.
If H is a homomorphism from P_j to X, then look at the two endpoints of the path, say B and B'. When B and B' are at distance j from each other, we'll say that P_j has property $\mathcal{P}$.

Pick $q_{P_i} = \frac{(k-1)!}{(k-i-1)!(k-1)^i}$ and $f = (k-1)/4$.

We need to check that $\mathcal{P}$ is weakly injective-like. The first bullet point of the definition of weakly injective-like (definition 9) is that property $\mathcal{P}$ holds for all homomorphisms of P_0. But a single vertex is always at distance 0 from itself so this is true.

Setting $i = 0$ we get $q_{P_0} = 1$ so the second bullet point is satisfied.

For the final bullet point, we are given a homomorphism $H : P_{i-1} \to X$ satisfying $\mathcal{P}$ and asked to count the number of ways extend it to some $H' : P_i \to X$ in order to continue satisfying property $\mathcal{P}$. Let the image of H consist of vertices $B_0, B_1, B_2, ..., B_{i-1}$, so now we are looking for some vertex B_i to extend the path, i.e. $B_i = B_{i-1} \cup \{x_i\} \backslash \{y_i\}$ with an edge between B_i and B_{i-1}. Given a specific choice of x_i, there are exactly $k-1$ choices of y_i that make (B_i, B_{i-1}) an edge, and so there are exactly $\frac{\deg(B_{i-1})}{k-1}$ choices for x_i. But not all of these choices satisfy property $\mathcal{P}$. How many of them do? Well, because H satisfies $\mathcal{P}$, we know that B_{i-1} is at distance $i-1$ from B_0 so we can write $B_{i-1} = B_0 \cup \{x_1, x_2, ..., x_{i-1}\} \backslash \{y_1, y_2, ..., y_{i-1}\}$. To make $B_i = B_{i-1} \cup \{x_i\} \backslash \{y_i\}$ be at distance i from B_0, it suffices that x_i is not any of the y_js and that y_i is not any of the x_js. So after excluding these cases, there are at least $\frac{\deg(B_{i-1})}{k-1} - (i-1)$ choices for x_i, followed by at least $k-i$ choices for y_i. The overall number of choices for H' is therefore at least $\frac{\deg(B_{i-1})(k-i)}{k-1} - \frac{(i-1)(k-i)}{k-1} \geq \frac{\deg(B_{i-1})(k-i)}{k-1} - \frac{k-1}{4}$.

This proves that $\mathcal{P}$ is weakly injective-like. So if Conjecture 1 is true, then we can apply it. It tells us that the number of paths in X joining two points at distance i from each other is at least $b \cdot [k(k-1)a/b]^i \cdot \prod_{j=1}^{i} \frac{k-j}{k-1} \cdot \left(1 - \left(\frac{\ln(b)b}{a}\right)\right)$, or to put it more neatly:

$$|\operatorname{Hom}_{\mathcal{P}}(P_i, X)| \geq \frac{(k-1)!}{(k-i-1)!} \left(\frac{ak}{b}\right)^i b \left(1 - O\left(\frac{b\ln(b)}{ka}\right)\right). \tag{3.5}$$

3.6.1 Upper bound on a as a function of b

Given any pair of vertices at distance i from each other, say B and
$B' = B \cup \{x_1, x_2, ..., x_i\} \backslash \{y_1, y_2, ..., y_i\}$ there are at most $i!^2$ paths of length i between them. This is because to get a path of length i, you need to pick some sequence of the x_js and y_js and you then follow the path by adding the next x_j and subtracting the

next y_j at each step; this then determines the path uniquely. So therefore the number of pairs of vertices at distance i from each other is at least $\frac{|\mathrm{Hom}_{\mathcal{P}}(T_i,X)|}{i!^2}$. So the total number of pairs, (which is equal to $b(b-1)$), is at least $\sum_{i=1}^{k-1}\frac{|\mathrm{Hom}_{\mathcal{P}}(T_i,X)|}{i!^2}$. Thus:

$$b(b-1) \geq \sum_{i=1}^{k-1}\left(\frac{ak}{b}\right)^i \frac{(k-1)!}{(k-i-1)!i!^2} b\left(1-O\left(\frac{b\ln(b)}{a}\right)\right)$$

$$b-1 \geq \left(\frac{ak}{b}\right)^{k-1}/(k-1)! + O\left(\left(\frac{ak}{b}\right)^{k-2}\ln(b)\right)$$

$$\frac{ak}{b} \leq (b(k-1)!)^{1/(k-1)} + O(\ln(b))$$

$$a \leq b^{k/(k-1)}\frac{(k-1)!^{1/(k-1)}}{k} + O(b\ln(b).)$$

This matches our canonical example where $a = \binom{c}{k}$ and $b = \binom{c}{k-1}$ to within $O(b\ln(b))$. Indeed, the canonical example has $\frac{ak}{b} = c-k+1$ and $b(k-1)! = c(c-1)...(c-k+2) = c^{k-1}+O(c^{k-2})$, so $(b(k-1)!)^{1/(k-1)} = c+O(1)$, therefore $\frac{ak}{b} = (b(k-1)!)^{1/(k-1)}+O(1)$ in our canonical example.

3.6.2 Using stability to gather information about our sets

This is similar to what we did in the case $k=3$. Suppose we have a valid configuration with b vertices and $a = b^{k/(k-1)}\frac{(k-1)!^{1/(k-1)}}{k}\left(1+O\left(\frac{\ln(b)}{b^{1/k-1}}\right)\right)$ edges. We will go through the proof to see what properties we can deduce of $\mathcal{A}$ and $\mathcal{B}$.

Nice pairs

We know from (3.5) that

$$|\mathrm{Hom}_{\mathcal{P}}(T_{k-1},X)| \geq (k-1)!\left(\frac{ak}{b}\right)^{k-1} b\left(1-O\left(\frac{b\ln(b)}{ka}\right)\right) = (k-1)!^2b^2\left(1-O\left(\frac{\ln(b)}{b^{1/(k-1)}}\right)\right).$$

We know that there are less than b^2 pairs of points at distance $k-1$. The maximum number of paths in $\mathrm{Hom}_{\mathcal{P}}(T_{k-1},X)$ joining such a pair is $(k-1)!^2$. Most pairs should have exactly $(k-1)!^2$ paths; we'll call this a nice pair. However, there might be some that have less. We'll say that there are ω pairs that are not nice. Then the number of paths of length $k-1$ is less than $(b^2-\omega)(k-1)!^2+\omega[(k-1)!^2-1] = b^2(k-1)!^2-\omega$. Combining this with our inequality for $\mathrm{Hom}_{\mathcal{P}}(T_{k-1},X)$, this tells us that $\omega \leq O\left(b^2\frac{\ln(b)}{b^{1/(k-1)}}\right)$. So almost all pairs at distance $k-1$ will be nice. The exceptions only make up a proportion

of $O\left(\frac{\ln(b)}{b^{1/(k-1)}}\right)$ of the total. We also get that $|\operatorname{Hom}_{\mathcal{P}}(T_{k-1}, X)| \leq (k-1)!^2 b^2$.

So given a vertex B, the average number of points B'' that create a nice pair with it is $b\left(1 - O\left(\frac{\ln(b)}{b^{1/(k-1)}}\right)\right)$. We want to know how many vertices are close to that number. To be more precise, lets say that there are $b(1-\delta)$ vertices B that have at least $b(1-\epsilon)$ vertices B'' at distance $k-1$ from it (where ϵ and δ will be defined later). Then the total number of pairs at distance $k-1$ is:

$$b^2\left(1 - O\left(\frac{\ln(b)}{b^{1/(k-1)}}\right)\right) \leq b(b - b\epsilon) + (b - b\delta)(b\epsilon).$$

Reordering the inequality gives:

$$\epsilon \leq \frac{1}{\delta} O\left(\frac{\ln(b)}{b^{1/(k-1)}}\right).$$

So if we set $\delta = b^{-1/(2k-2)}$ and $\epsilon = O\left(\frac{\ln(b)}{b^{1/(2k-2)}}\right)$, then this works. Therefore there are $b\left(1 - b^{-1/(2k-2)}\right)$ vertices B that are part of at least $b\left(1 - O\left(\frac{\ln(b)}{b^{1/(2k-2)}}\right)\right)$ nice pairs. We'll call this set of vertices $\mathcal{B}'$.

Low degree

We know that the sum of the degrees of all the vertices is ka. Therefore the sum of the degrees of vertices in $\mathcal{B}'$ is at most ka. So the average degree of an element of $\mathcal{B}'$ is at most $\frac{ka}{b\left(1-b^{-1/(2k-2)}\right)} \leq (b(k-1)!)^{1/(k-1)}(1+\epsilon)$.

Therefore there exists a vertex B of $\mathcal{B}'$ with degree less than $(b(k-1)!)^{1/(k-1)}(1+\epsilon)$.

Colouring the edges incident to B

We know we have a vertex B that is part of at least $b(1-\epsilon)$ nice pairs, and moreover, it has degree $d \leq (b(k-1)!)^{1/(k-1)}(1+\epsilon)$. We'll denote the set of edges incident to B by $\mathcal{G} = \{B \cup \{x_1\}, B \cup \{x_2\}, ..., B \cup \{x_d\}\}$.

Each of the edges in $\mathcal{G}$ has $k-1$ other vertices incident to it: $B \cup \{x_i\} \backslash \{y_{i,1}\}$, $B \cup \{x_i\} \backslash \{y_{i,2}\}$, ..., $B \cup \{x_i\} \backslash \{y_{i,k-1}\}$. We'll also colour $\mathcal{G}$ by giving colour: $\{y_{i,1}, y_{i,2}, ..., y_{i,k-1}\}$ to the edge $B \cup \{x_i\}$.

Now for every nice pair (B, B''), there are $(k-1)!^2$ paths between them. B'' is at

distance $k-1$ from B so write $B'' = B \cup \{t_1, t_2, ..., t_{k-1}\} \backslash \{w_1, w_2, ..., w_{k-1}\}$. Now there exist all $(k-1)!^2$ possible paths between B and B'', which means that $B \cup \{t_1\}$, $B \cup \{t_2\}$, ..., $B \cup \{t_{k-1}\}$ are all in $\mathcal{G}$. Therefore , each t_i is equal to some x_j. Furthermore, all these edges have the same colour: $\{w_1, w_2, ..., w_{k-1}\}$. So therefore we know that for every nice pair (B, B''), there exists a corresponding set of $k-1$ elements of $\mathcal{G}$ that all have the same colour. Furthermore, given such a monochromatic set of $k-1$ elements of $\mathcal{G}$, there is at most one B'' that they correspond to. So the number of nice pairs containing B is at most the number of monochromatic $(k-1)$-sets of $\mathcal{G}$.

We'll say that the largest colour class of $\mathcal{G}$ has size $d(1-\alpha)$. Then the maximum number of monochromatic $(k-1)$-sets is $\binom{d\alpha}{k-1} + \binom{d(1-\alpha)}{k-1} = \frac{d^{k-1}}{(k-1)!}(\alpha^{k-1} + (1-\alpha)^{k-1} - O(1/d))$. We know that this number is at least $b(1-\epsilon)$ and so we plug in $d \leq (b(k-1)!)^{1/(k-1)}(1+\epsilon)$ to get:

$$\begin{aligned}
\frac{b(k-1)!}{(k-1)!}(\alpha^{k-1} + (1-\alpha)^{k-1} - O(1/d))(1+\epsilon)^{k-1} &\geq b(1-\epsilon) \\
(\alpha^{k-1} + (1-\alpha)^{k-1} - O(1/d)) &\geq \frac{1-\epsilon}{(1+\epsilon)^{k-1}} \\
\alpha &\leq \frac{k}{k-1}\epsilon + O(\epsilon^2) + O(1/d).
\end{aligned}$$

Therefore, we know that there is a very large colour class of size $d(1-\frac{k\epsilon}{k-1}+O(\epsilon^2)+O(1/d))$, comprising nearly all elements of $\mathcal{G}$. We'll say its colour is $\{z_1, z_2, ..., z_{k-1}\}$.

A nice hypergraph

We want to know how many elements of $\mathcal{B}'$ are connected to our large colour class. The maximum number of them that we don't use in our large colour class is $\binom{\alpha}{k-1} \leq \left(\frac{k\epsilon}{k-1}\right)^{k-1} \frac{d^{k-1}}{(k-1)!} \leq \left(\frac{k\epsilon(1+\epsilon)}{k-1}\right)^{k-1} b$. Therefore the number of vertices of $\mathcal{B}'$ that are connected to our large colour class is at least $b\left(1-\epsilon-\left(\frac{k\epsilon(1+\epsilon)}{k-1}\right)^{k-1}\right)$, so that is nearly all points.

Now notice that every one of these vertices of $\mathcal{B}'$ that is connected to our large colour class contains $B\backslash\{z_1, z_2, ..., z_{k-1}\}$ because that is the only way to connect it to edges in $\mathcal{G}$ of that colour. We set $S = B\backslash\{z_1, z_2, ..., z_{k-1}\}$ and we end up with a nice hypergraph that comprises nearly all the vertices of $\mathcal{B}$. So what we have is:

Lemma 3. *There exists a set $S \in \mathbb{N}^{(r-k)}$ such that S is a subset of $b(1-o(1))$ elements of $\mathcal{B}$.*

We'll define our nice hypergraph $\mathcal{D}$ to consist of all the vertices and edges that contain S as a subset.

3.6.3 Using classical Kruskal Katona to improve the bound further

At this point, we know that we have a large nice hypergraph of vertices $\mathcal{D}$, all of which contain S as a subset. We'll say that there are λ vertices in $\mathcal{B}\backslash\mathcal{D}$. We know that $\lambda = o(b)$. Our aim in this section is to bound λ by a constant. How many edges can we have in our graph? To count them, we'll separate them into 3 cases:

• The edges that are entirely contained within $\mathcal{D}$. We can apply the classical version of the Kruskal Katona Theorem to get an upper bound. To get that bound, we need to write $|\mathcal{D}| = b - \lambda$ in the form $\binom{d_{k-1}}{k-1} + \binom{d_{k-2}}{k-2} + ... + \binom{d_1}{1}$. Then the maximum number of edges is $\binom{d_{k-1}}{k} + \binom{d_{k-2}}{k-1} + ... + \binom{d_1}{2}$.

• The edges that are entirely contained within $\mathcal{B}\backslash\mathcal{D}$. For these, we can just apply our formula to say that there are at most $\lambda^{k/(k-1)} \frac{(k-1)!^{1/(k-1)}}{k} \left(1 + O\left(\frac{\ln(\lambda)}{\lambda^{1/(k-1)}}\right)\right)$ of them.

• The edges that are incident to both $\mathcal{D}$ and $\mathcal{B}\backslash\mathcal{D}$. Since these edges are incident to some vertex in our nice hypergraph $\mathcal{D}$, that vertex has to contain S as a subset, therefore the edge also has to contain S. Now for every vertex B in $\mathcal{B}\backslash\mathcal{D}$, $S \not\subset B$, so the only edge that can connect B to $\mathcal{D}$ has to be $B \cup S$. In particular, it is unique. Therefore the number of edges that fall under this case is at most λ.

If we add up everything, we get that the maximal number of edges is:
$\binom{d_{k-1}}{k} + \binom{d_{k-2}}{k-1} + ... + \binom{d_1}{2} + O(\lambda^{k/(k-1)})$, where $\lambda = b - \left[\binom{d_{k-1}}{k-1} + \binom{d_{k-2}}{k-2} + ... + \binom{d_1}{1}\right]$.

We want to compare this to what we would get with our hypothesis (which states that $\lambda = 0$ is optimal). For this, you would write $b = \binom{c_{k-1}}{k-1} + \binom{c_{k-2}}{k-2} + ... + \binom{c_1}{1}$, and then the number of edges would be: $\binom{c_{k-1}}{k} + \binom{c_{k-2}}{k-1} + ... + \binom{c_1}{2}$. So if we did have a counter-example to our hypothesis, we would have:

$$\left[\binom{c_{k-1}}{k} + \binom{c_{k-2}}{k-1} + ... + \binom{c_1}{2}\right] - \left[\binom{d_{k-1}}{k} + \binom{d_{k-2}}{k-1} + ... + \binom{d_1}{2}\right] = O(\lambda^{k/(k-1)})$$

$$\text{where } \lambda = \left[\binom{c_{k-1}}{k-1} + \binom{c_{k-2}}{k-2} + ... + \binom{c_1}{1}\right] - \left[\binom{d_{k-1}}{k-1} + \binom{d_{k-2}}{k-2} + ... + \binom{d_1}{1}\right].$$

We split into several cases depending on what $c_{k-1} - d_{k-1}$ is.

Case 1: $c_{k-1} - d_{k-1} \geq 2$

Then $\lambda \leq \binom{c_{k-1}+1}{k-1} - \binom{d_{k-1}}{k-1} \leq (c_{k-1}+1-d_{k-1})\binom{c_{k-1}}{k-2} = (c_{k-1}+1-d_{k-1})O((c_{k-1})^{k-2})$. Therefore the right hand side of the inequality is at most $(c_{k-1}+1-d_{k-1})^{k/(k-1)}O\left((c_{k-1})^{k(k-2)/(k-1)}\right)$.

Meanwhile, the left hand side of the inequality is at least $\binom{c_{k-1}}{k} - \binom{d_{k-1}+1}{k} \geq (c_{k-1} - d_{k-1} - 1)\binom{d_{k-1}+1}{k-1} = (c_{k-1} - d_{k-1} - 1)\Omega(c_{k-1}^{k-1})$.

Notice that $(c_{k-1} - d_{k-1} - 1) \geq 1$, so to get the inequality to hold, we must have $\frac{(c_{k-1}-d_{k-1}+1)^{k/(k-1)}}{c_{k-1}-d_{k-1}-1} > \Omega((c_{k-1})^{1/(k-1)})$, so therefore $c_{k-1} - d_{k-1} = \Omega(c_{k-1})$. But we know that $\lambda = o(b)$ so $(c_{k-1})^{k-1} - (d_{k-1})^{k-1} = o((c_{k-1})^{k-1})$ so $c_{k-1} - d_{k-1} = o(c_{k-1})$. This is a contradiction, so the inequality never holds when $c_{k-1} - d_{k-1} \geq 2$, so there are no counter-examples of this type (as long as c_{k-1} is large).

Case 2: $c_{k-1} - d_{k-1} = 1$

We substitute $d_{k-1} = c_{k-1} - 1$ into the inequality to get:

$$\binom{c_{k-1}-1}{k-1} + \left[\binom{c_{k-2}}{k-1} + ... + \binom{c_1}{2}\right] - \left[\binom{d_{k-2}}{k-1} + ... + \binom{d_1}{2}\right] < O(\lambda^{k/(k-1)})$$

$$\text{where } \lambda = \binom{c_{k-1}-1}{k-2} + \left[\binom{c_{k-2}}{k-2} + ... + \binom{c_1}{1}\right] - \left[\binom{d_{k-2}}{k-2} + ... + \binom{d_1}{1}\right].$$

Then $\lambda \leq \binom{c_{k-1}}{k-2} = O((c_{k-1})^{k-2})$. Therefore the right hand side is at most $O((c_{k-1})^{k(k-2)/(k-1)})$.

Meanwhile, the left hand side of the inequality is at least:
$\binom{c_{k-1}-1}{k-1} - \binom{d_{k-2}+1}{k-1} =$
$\frac{(c_{k-1})^{k-1}}{(k-1)!} - O((c_{k-1})^{k-2}) - \frac{(d_{k-2})^{k-1}}{(k-1)!} + O((d_{k-2})^{k-2})$. The only way to get this to be smaller

than the right hand side is to have the $\frac{(c_{k-1})^{k-1}}{(k-1)!} - \frac{(d_{k-2})^{k-1}}{(k-1)!} = O((c_{k-1})^{k-1-1/(k-1)})$ so $c_{k-1} - d_{k-2} = O((c_{k-1})^{1-1/(k-1)})) = o(c_{k-1})$.

Now we try again except this time we can use the information that $c_{k-1} - d_{k-2} = o(c_{k-1})$. We have $\lambda \leq \binom{c_{k-1}-1}{k-2} + \binom{c_{k-2}+1}{k-2} - \binom{d_{k-2}}{k-2} \leq (c_{k-1} - d_{k-2} - 1)\binom{c_{k-1}-2}{k-3} + \binom{c_{k-2}+1}{k-2}$. We can use Jensen's inequality to deduce that the right hand side of the inequality is at most: $(c_{k-1} - d_{k-2} - 1)^{k/(k-1)} O((c_{k-1})^{k(k-3)/(k-1)}) + O((c_{k-2})^{k(k-2)/(k-1)})$.

Meanwhile, the left hand side of the inequality is at least: $\binom{c_{k-1}-1}{k-1} + \binom{c_{k-2}}{k-1} - \binom{d_{k-2}+1}{k-1} \geq (c_{k-1} - d_{k-2} - 2)\binom{d_{k-2}+1}{k-2} + \binom{c_{k-2}}{k-1} = (c_{k-1} - d_{k-2} - 2)\Omega((c_{k-1})^{k-2}) + \Omega((c_{k-2})^{k-1})$. The only way this is smaller than the right hand side is if $(c_{k-1} - d_{k-2} - 2)\Omega((c_{k-1})^{k-2}) < (c_{k-1} - d_{k-2} - 1)^{k/(k-1)} O((c_{k-1})^{k(k-3)/(k-1)})$. This implies that $\frac{c_{k-1}-d_{k-2}-2}{(c_{k-1}-d_{k-2}-1)^{k/(k-1)}} < O((c_{k-1})^{-2/(k-1)})$.

There are two solutions to this: either $c_{k-1} - d_{k-2} - 2 = 0$, or $c_{k-1} - d_{k-2} = \Omega((c_{k-1})^2)$. But the second solution is clearly impossible, so the only possibility is $d_{k-2} = c_{k-1} - 2$. We substitute this back into the inequality and we get:

$$\binom{c_{k-1}-2}{k-2} + \left[\binom{c_{k-2}}{k-1} + ... + \binom{c_1}{2}\right] - \left[\binom{d_{k-3}}{k-2} + ... + \binom{d_1}{2}\right] < O(\lambda^{k/(k-1)})$$

$$\text{where } \lambda = \binom{c_{k-1}-2}{k-3} + \left[\binom{c_{k-2}}{k-2} + ... + \binom{c_1}{1}\right] - \left[\binom{d_{k-3}}{k-3} + ... + \binom{d_1}{1}\right].$$

This looks remarkably like the inequality we had at the start of this case. In fact, we can repeat the argument with a few changes to prove that $d_{k-3} = c_{k-1} - 3$. Then we can continue using the same argument to prove $d_{k-4} = c_{k-1} - 4$, ... all the way until $d_1 = c_{k-1} - (k-1)$. At this point, we're left with:

$$(c_{k-1} - k + 1) + \left[\binom{c_{k-2}}{k-1} + ... + \binom{c_1}{2}\right] < O(\lambda^{k/(k-1)})$$

$$\text{where } \lambda = 1 + \left[\binom{c_{k-2}}{k-2} + ... + \binom{c_1}{1}\right].$$

We need $\lambda^{k/(k-1)} \geq \Omega(c_{k-1})$ so $\binom{c_{k-2}+1}{k-2} \geq \lambda \geq \Omega((c_{k-1})^{(k-1)/k})$ so $c_{k-2} \geq \Omega((c_{k-1})^{(k-1)/k/(k-2)})$. Because $\frac{(k-1)^2}{k(k-2)} > 1$, the dominant term on the left hand side of

the inequality is $\binom{c_{k-2}}{k-1}$. Therefore we end up with $\Omega((c_{k-2})^{k-1}) < O((c_{k-2})^{k(k-2)/(k-1)})$ which is impossible for c_{k-2} large enough. $c_{k-2} \geq \Omega((c_{k-1})^{(k-1)/k/(k-2)})$ so it's also impossible when c_{k-1} is large enough. So the inequality never holds and there are no large counter-examples of this type.

Case 3: $c_{k-1} - d_{k-1} = 0$

We substitute $d_{k-1} = c_{k-1}$ into the inequality to get:

$$\left[\binom{c_{k-2}}{k-1} + ... + \binom{c_1}{2}\right] - \left[\binom{d_{k-2}}{k-1} + ... + \binom{d_1}{2}\right] < O(\lambda^{k/(k-1)})$$

$$\text{where } \lambda = \left[\binom{c_{k-2}}{k-2} + ... + \binom{c_1}{1}\right] - \left[\binom{d_{k-2}}{k-2} + ... + \binom{d_1}{1}\right].$$

But $O(\lambda^{k/(k-1)}) < O(\lambda^{(k-1)/(k-2)})$, so we get:

$$\left[\binom{c_{k-2}}{k-1} + ... + \binom{c_1}{2}\right] - \left[\binom{d_{k-2}}{k-1} + ... + \binom{d_1}{2}\right] < O(\lambda^{(k-1)/(k-2)})$$

$$\text{where } \lambda = \left[\binom{c_{k-2}}{k-2} + ... + \binom{c_1}{1}\right] - \left[\binom{d_{k-2}}{k-2} + ... + \binom{d_1}{1}\right].$$

This new inequality is identical to our original except that we have $k-1$ instead of k. If c_{k-2} is large enough, then we can repeat our argument and either go to cases 1 and 2 and prove no counter-example exists, or go back through case 3 and reduce k again (more formally, there exists a constant μ depending only on k such that if for any i, $d_i < c_i > \mu$, then case 1 or 2 applies and there is no counter-example possible). If all of c_{k-2}, c_{k-3}, ..., c_1 are larger than μ, then we get $c_{k-1} = d_{k-1}$, $c_{k-2} = d_{k-2}$, ... , $c_1 = d_1$, which implies that $b = b - \lambda$ so $\lambda = 0$.

This is exactly Theorem 4: there is some constant μ depending only on k such that if $b = \left[\binom{c_{k-1}}{k-1} + ... + \binom{c_1}{1}\right]$, for some $c_{k-1} > c_{k-2} > ... > c_1 > \mu$, then the maximum value for a is exactly:

$$\left[\binom{c_{k-1}}{k} + ... + \binom{c_1}{2}\right] = f(r, k, b).$$

The only cases that aren't covered by Theorem, is if there exists some i such that c_i is

smaller than μ. Without loss of generality, pick i to be the largest such. Then we know that $c_j = d_j$ for any $j > i$. Then we get:

$$\lambda \leq \left[\binom{\mu}{i}\binom{\mu-1}{i-1} + ... + \binom{\mu-i+1}{1} \right] - [0].$$

So λ is in fact bounded by a constant, which implies the following lemma:

Lemma 4. *There is a constant λ_{max} depending only on k such that there exists a subset $\mathcal{D}$ of $\mathcal{B}$ of size at least $b - \lambda_{max}$ and a set $S \in \mathbb{N}^{(r-k)}$ such that S is a subset of every element of $\mathcal{D}$.*

And now Theorem 5 is just an easy corollary of this, since we have a bounded number of vertices that aren't in our nice hypergraph, these vertices can only form a bounded number of extra edges, therefore there is a constant τ depending only on k such that if $b = \left[\binom{c_{k-1}}{k-1} + ... + \binom{c_1}{1} \right]$, for some $c_{k-1} > c_{k-2} > ... > c_1$, then the maximum value for a is between:

$$\left[\binom{c_{k-1}}{k} + ... + \binom{c_1}{2} \right] \leq f(r,k,b) \leq \left[\binom{c_{k-1}}{k} + ... + \binom{c_1}{2} \right] + \tau.$$

Chapter 4

Rational Exponents for Turán Hypergraph Problems

4.1 Introduction

Definition 14. *Given an integer $k \geq 2$ and a family $\mathcal{F}$ of k-hypergraphs, $ex(n, \mathcal{F})$ is defined to be the maximum number of edges across all k-hypergraphs that have n vertices and do not contain any element of $\mathcal{F}$ as a subgraph.*

In general, $ex(n, \mathcal{F})$ can be anything from 0 (as in the case $\mathcal{F} = \{E\}$, where E is just a single edge) to $\binom{n}{k}$ (as in the case where $\mathcal{F}$ is empty).

In this chapter, we will extend Bukh and Conlon's [8] result to hypergraphs, ie, we will prove that for every rational r between 1 and k, there exists a finite family of k-hypergraphs $\mathcal{F}$ with $ex(n, \mathcal{F}) = \Theta(n^{k-r})$. This is also an improvement on Frankl's [13] result since we now have a family of k-hypergraphs for all $k \geq r$, instead of for just one specific k. This is also of interest as an infinite family of k-hypergraphs for which the answer to the Turán problem is known. Note that in the literature, this is commonly formulated as $ex(n, \mathcal{F}) = \Theta(n^r)$ for some $k \geq r$ instead of $\Theta(n^{k-r})$ as we do. The reason we prefer to exchange r and $k - r$ is because it will make the calculations easier.

To prove this we will use similar methods as Bukh and Conlon [8], both in the construction of $\mathcal{F}$ and for the lower bound. For the upper bound, we will use Theorem 2 from chapter 2.

We will at first only consider the case where $0 \leq r < 1$:

Theorem 6. *Given an integer k and a rational r, $0 \leq r < 1$, there exists some finite collection of k-hypergraphs $\mathcal{F}$ such that $ex(n, \mathcal{F}) = \Theta(n^{k-r})$.*

Our section 4.2 deals with the construction of the family of hypergraphs $\mathcal{F}$ that will solve Theorem 6. They will be hypergraph versions of the graphs from [8].

In section 4.3, we prove the lower bound, i.e. that $ex(n, \mathcal{F}) \geq \Theta(n^{k-r})$. This involves constructing a hypergraph with n vertices and $\Theta(n^{k-\frac{a}{b}})$ edges but that does not contain any copy of any hypergraph from $\mathcal{F}$. The proof is again adapted from [8].

In section 4.4, we prove the upper bound, i.e. that $ex(n, \mathcal{F}) \leq \Theta(n^{k-r})$. However, unlike in the first two sections, the proof from [8] cannot be easily extended to hypergraphs. We instead use Theorem 2 that we proved in chapter 2. When n is a sufficiently large integer, this allows us to find some copy of an element of $\mathcal{F}$ in any hypergraph X with n vertices and with at least $\Theta(n^{k-r})$ edges, thereby proving the upper bound.

In our final section, we consider what happens for other rs. We first extend the result from $0 \leq r < 1$ to $0 \leq r \leq k-1$:

Theorem 7. *Given an integer k and a rational r, $0 \leq r < k-1$, there exists some collection of k-hypergraphs $\mathcal{F}$ such that $ex(n, \mathcal{F}) = \Theta(n^{k-r})$.*

Observation: The case where $k-1 < r < k$ is impossible. This is a corollary of the Sunflower Lemma [11], which involves hypergraphs called *sunflower*s. A sunflower is a k-hypergraph which contains a set of between 0 and $k-1$ points, called the *kernel*, such that any two edges of the sunflower intersect in exactly the kernel. The Sunflower Lemma states that whenever F is a collection of k-hypergraphs such that for all $0 \leq i \leq k-1$, F contains a sunflower with kernel size i, then $ex(n, F)$ is bounded by a constant (independent of n). We shall provide more details in the final section.

Algebraic Geometry

The proof will use some algebraic geometry. What follows in this section is a brief overview of the results we will use. See [14] for more information and proofs.

Definition 15. *(1.1.2 in [14]) Given an algebraically closed field $\mathbb{F}$, an* affine algebraic variety *V (often shorted to just variety) over $\mathbb{F}$ is a set of the form:*
$V = \{(x_1, x_2, ..., x_n) \in \mathbb{F}^n \mid P_1(x_1, x_2, ..., x_n) = P_2(x_1, ..., x_n) = ... = P_m(x_1, ..., x_n) = 0\}$,

where $P_1, P_2, ..., P_m$ are polynomials over $\mathbb{F}$ with n variables.

Lemma 5. *(1.1.4 and 1.1.5 in [14]) If U and V are varieties over $\mathbb{F}$, then $U \cap V$ and $U \cup V$ are also varieties over $\mathbb{F}$.*

Definition 16. *(1.1.10 in [14]) Given a variety V over $\mathbb{F}$, we say that V is* reducible *if there exist varieties $U, U' \subsetneq V$ such that $V = U \cup U'$. If V is not reducible, we say it is* irreducible*.*

Lemma 6. *(1.1.12a in [14]) A variety V can be decomposed uniquely (up to ordering) into maximal irreducible components: $V = U_1 \cup U_2 \cup ... \cup U_k$ where the U_i are all irreducible and such that for all i, j, $U_i \not\subset U_j$.*
That means that if we have two such decompositions $V = \bigcup_{i=1}^{k} U_i = \bigcup_{j=1}^{l} U'_j$, then $k = l$ and for every $1 \leq i \leq k$, there exists some $1 \leq j \leq l$ such that $U_i = U'_j$ and vice-versa.

Furthermore, the number of components in such a decomposition is bounded above by d^m where m is the number of polynomials that generate the variety, and d is their maximum degree.

Definition 17. *(1.2.15 to 1.2.17 in [14]) Given a non-empty irreducible variety V over $\mathbb{F}$, its* dimension *δ is the length of the longest sequence: $V = V_\delta \supsetneq V_{\delta-1} \supsetneq V_{\delta-2} \supsetneq ... \supsetneq V_0 \supsetneq \emptyset$, where every V_i is irreducible. This is well defined for every non-empty irreducible variety.*
When V is reducible, we say its dimension is the largest dimension of its irreducible components.

It is fairly easy to see that a finite set of points has dimension 0, the space $\mathbb{F}^n$ has dimension n, and that when V is a non-empty variety generated by k polynomials in n variables: $P_1, P_2, ..., P_m \in \mathbb{F}[X_1, X_2, ..., X_n]$, then V has dimension at least $n - m$.

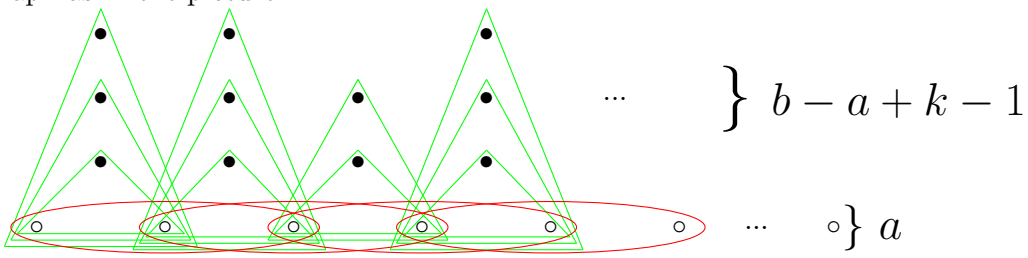
Although we require $\mathbb{F}$ to be algebraically closed for the theory to work, most practical applications involve fields that are not algebraically closed. However, this isn't a problem because if $\mathbb{F}'$ is an arbitrary field, then it has an algebraic closure $\overline{\mathbb{F}'}$. We can then use properties of algebraic varieties over $\overline{\mathbb{F}'}$ to say things about the corresponding set over $\mathbb{F}'$:

Definition 18. *Given a variety V over an algebraically closed field $\mathbb{F}$ and a subfield $\mathbb{F}' \subseteq \mathbb{F}$ (which might not be algebraically closed), the* $\mathbb{F}'$-rational points *of the variety, denoted by $V(\mathbb{F}')$, are defined to be the points of V that can be written using elements of $\mathbb{F}'$, i.e.: $V(\mathbb{F}') = V \cap \mathbb{F}'^n$.*

Theorem 8. *(Lang-Weil bound) [20] Let $\mathbb{F}_p$ be the finite field of order p, where p is a power of a prime. Let V be an irreducible variety of dimension δ over $\overline{\mathbb{F}_p}$. Then $V(\mathbb{F}_p)$ is either empty or has $|V(\mathbb{F}_p)| = p^{\delta}(1 + O(p^{-1/2}))$*

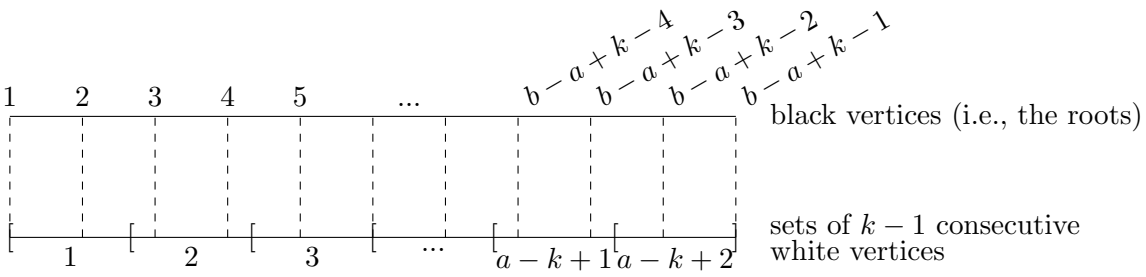
4.2 The set of hypergraphs

Suppose we are given an integer k and some rational $0 \leq r = \frac{a}{b} < 1$ Since r is a rational smaller than 1, $b > a$ and both are positive integers. By multiplying a and b by some constant, we can assume without loss of generality that $b \geq a - k + 3$. Now consider the hypergraph as in the picture:



Example of the hypergraph $\mathcal{T}$ in the case $k = 3$

It is essentially a hypergraph version of the graph from [8]. It is comprised of an ordered set of a vertices (in white) with edges (in red) being sets of k vertices in a row. We add to this $b-a+k-1$ vertices (in black) and for each one, an edge (in green) connecting it to $k-1$ vertices in a row. This makes the total number of edges to be b. These black vertices are as evenly spaced as possible (see picture below). Formally, the ith black vertex is connected to the $\lfloor 1 + \frac{(i-1)(a-k+2)}{b-a+k-2} \rfloor$th $(k-1)$-set of consecutive white vertices. There is one exception, and that is the last (i.e. $(b-a+k-1)$th) black vertex is connected to the last (i.e.: $(a-k+2)$th) consecutive set of white vertices, not, as the formula suggests, the $(a-k+3)$th, because that one doesn't exist. We call the vertex-set of this hypergraph T, the subset of black vertices R and call these black vertices the *roots* of $\mathcal{T}$.



An example of how the roots are connected to the sets of $k-1$ consecutive non-roots

In this picture for example, the second $(k-1)$-set of non-roots is connected to both the 3rd and 4th root but no others. When a root lands exactly on a border, it gets connected to the $(k-1)$-set corresponding to the interval on its right EXCEPT for the very last one, which gets connected to the $(k-1)$-set on its left (because there is nothing to the right)

4.2.1 $\mathcal{T}$ is balanced

Definition 19. *Given a set S of non-roots, define $\epsilon(S)$ to be the number of edges that contain a point of S.*

Definition 20. *A rooted k-hypergraph $\mathcal{U}$ with vertex set U and set of roots R is* balanced *if for any subset $S \subset U - R$, we have:*

$$\frac{\epsilon(S)}{|S|} \geq \frac{\epsilon(U-R)}{|U-R|}.$$

Notice that in the case of our hypergraph $\mathcal{T}$ (whose vertex set is T), we have $\epsilon(T-R)$ is the total number of edges in the hypergraph, i.e. $\epsilon(T-R) = b$.

Lemma 7. *The hypergraph $\mathcal{T}$ defined above is balanced.*

Proof: First of all, if every edge of $\mathcal{T}$ contains an element of S, then the result is trivial since $|S| \leq |T-R|$. Without loss of generality, we can therefore assume that there is at least one edge that doesn't contain any element of S, which means there is a section of length at least $k-1$ that does not contain any element of S. Call this a *hole*. We also separate S into a sequence of *blocks*, by which we mean a maximal sequence of elements of S with no gaps between them.

Suppose we have a block directly to the left of a hole and also suppose that it does not contain the leftmost vertex of T. Call this block B. What happens if we shift

the entire block to the left? Because the black vertices (roots) are evenly distributed, the number of green edges (edges containing roots) adjacent to B varies by at most 1. The number of red edges (edges not containing roots) containing a point of B stays the same unless we are reaching the left edge of $\mathcal{T}$, in which case it goes down. If we do not reach the left side of T, then that means there is another block in the way. In this case, the edges containing points of that block and the edges containing points of B will start to coincide. Regardless of which case we are in, when we do this step, the number of red edges containing elements of S goes down by at least 1, while the number of green edges changes by at most 1; therefore, the overall number of edges containing elements of S goes down or stays constant, while $|S|$ stays constant. Therefore we can assume without loss of generality that this step has been completed.

By repeating this step multiple times, we can move blocks left until they merge with other blocks, and then continue moving the bigger blocks until we eventually have everything to the left of the hole is in one big block as left as it can go. By a similar argument, everything to the right of the hole is in one big block as far right as it goes. Say the big block on the left has size x and the one on the right has size y.

If the total number of vertices in the left big block is x then we get x red edges. The green edges we get are those that connect to the first x $(k-1)$-sets. Recall from the definition, that the ith green edge connects to the $\lfloor 1+\frac{(i-1)(a-k+2)}{b-a+k-2}\rfloor$th $(k-1)$-set. Therefore the number of green edges is the maximal i such that $\lfloor 1+\frac{(i-1)(a-k+2)}{b-a+k-2}\rfloor \leq x$, i.e. s.t. $\frac{(i-1)(a-k+2)}{b-a+k-2} < x$, i.e. $i = \lceil \frac{x(b-a+k-2)}{a-k+2} \rceil$. Similarly, we can calculate the number of red edges in the right big block as y and the number of green edges in it as $\lfloor \frac{y(b-a+k-2)}{a-k+2} \rfloor + 1$. Therefore $\epsilon(S)$ is at least $|S| + \lceil \frac{|S|(b-a+k-2)}{a-k+2} \rceil = \lceil |S| \frac{b}{a-k+2} \rceil \geq |S| \frac{b}{a}$. Thus, $\mathcal{T}$ is indeed a balanced rooted hypergraph.

□

Definition 21. *$\mathcal{T}^{\leq s}$, the* sth power of the rooted hypergraph *$\mathcal{T}$, is defined to be the set of all k-hypergraphs that are formed by taking the union of s copies of $\mathcal{T}$ such that all the copies agree on the roots (that is to say, the ith root of the uth copy is the same as the ith root of the vth copy for all i,u and v). For the non-roots (i.e., the s copies of each non-root), any disposition is allowed: they can be distinct, they can coincide with each other, or they can even coincide with different non-roots from other copies of $\mathcal{T}$.*

Definition 22. *$\mathcal{T}^s = \mathcal{T}^{\leq s} \backslash \mathcal{T}^{\leq s-1}$.*

Lemma 8. *For any hypergraph H in $\mathcal{T}^s$, the number of edges in H is at least $(|H|-|R|)\frac{b}{a}$.*

Proof: We prove this lemma by induction on s. The case $s = 1$ is trivial since then $H = \mathcal{T}$.

Given $H \in \mathcal{T}^s$, we can write $v(H)$ as $v(H') \cup S$ where H' is in $\mathcal{T}^{s-1}$ and S is all the extra vertices from the sth copy of $\mathcal{T}$ that aren't already included in H. We can consider S as a subset of $T - R$. Since $\mathcal{T}$ is balanced, we have that the number of edges containing an element of S is at least $|S|\frac{b}{a}$. By induction, the number of edges in H' is at least $(|H'| - |R|)\frac{b}{a}$. Therefore the total number of edges in H is at least $(|S| + |H'| - |R|)\frac{b}{a} = (|H| - |R|)\frac{b}{a}$.

Therefore by induction, we have proved that H has at least $|H - R|\frac{b}{a}$ edges.

□

The set of hypergraphs $\mathcal{F}$ we will take to prove Theorem 6 is $\mathcal{T}^p$ for $p = 2(b^2(b - a + k - 1) + ab + b - 1)^{ab}$ and a and b are such that $r = \frac{a}{b}$.

4.3 The lower bound

To prove that $ex(n, \mathcal{F}) \geq \Theta(n^{k-r})$, we need to construct a hypergraph G with at least $\Theta(n^{k-r})$ edges but without any copies of $\mathcal{F}$. The hypergraph we will take will also be a hypergraph version of the graph from [8].

Denote $s = b(b - a + k - 1) + a + 1$, $d = bs - 1 = b^2(b - a + k - 1) + ab + b - 1$. Notice that now $p = 2d^{ab}$. Then pick a sufficiently large prime q.

The set of vertices of G is the disjoint union of k copies of $\mathbb{F}_q^b$: $\mathbb{F}_q^b \sqcup \mathbb{F}_q^b \sqcup ... \sqcup \mathbb{F}_q^b$. Also pick uniformly independently at random a polynomials in k variables of degree at most d: $f_1, f_2, ..., f_a$: $\mathbb{F}_q^b \times \mathbb{F}_q^b \times \mathbb{F}_q^b \times ... \times \mathbb{F}_q^b \to \mathbb{F}_q$ (there are k copies of $\mathbb{F}_q^b$). [Note: picking a polynomial of degree at most d at random here means that for every coefficient of degree $\leq d$, pick an element of $\mathbb{F}_q$ uniformly at random and independently of the others.] The edges of G are defined to be $(x_1, x_2, ..., x_k)$ such that $f_i(x_1, x_2, ..., x_k) = 0$ for all i.

Thus G is k-partite and has $kq^b = N$ vertices. The edges of G are equivalent to the rational points of the variety $V(\mathbb{F}_q)$, defined by a polynomials: $f_1, f_2, ..., f_a$. Now $V(\mathbb{F}_q)$ is either empty or contains some non-empty irreducible variety of same dimension as it. By the Lang-Weil bound [20], this irreducible variety has size at least $(1 - \Theta(q^{-1/2}) \cdot q^{dim(V)}$. Since we have only a polynomials defining the variety, we have $dim(V) \geq bk - a$ (unless it's empty). Therefore, either there are 0 edges in G, or there are at least $\Theta(q^{bk-a}) = \Theta(N^{k-\frac{a}{b}})$ edges in G, no matter which f_is we choose.

Probability that G is empty

Suppose we have already picked all the non-constant coefficients of all the f_is. Pick some points $(x_1, x_2, ..., x_k)$ arbitrarily. Then for each f_i, there is exactly one value for the constant coefficient that makes $f_i(x_1, ..., x_k) = 0$. The probability we pick it is $1/q$. Multiplying these together, the probability we pick exactly the right value for every f_i is $1/q^a$ because we picked the functions independently of each other. Therefore G contains $(x_1, x_2, ..., x_n)$ (and in particular, is non-empty) with probability at least $1/q^a$. For the next parts, we'll only be considering the case where G is indeed non-empty.

Claim: This hypergraph is $\mathcal{T}^p$-free with positive probability.

Proof:

Given a copy A of a hypergraph $H \in \mathcal{T}^{\leq s}$ in G, we know it has an ordered set of $b - a + k - 1$ roots. We'll call this ordered set $r(A) = (w_1, w_2, ..., w_{b-a+k-1}) = \mathbf{w}$.

Before finding a suitable $A \in \mathcal{T}^p$, we'll start by picking out a potential candidate for $r(A)$. This means we are arbitrarily picking an ordered set of $b - a + k - 1$ vertices: $(w_1, w_2, ..., w_{b-a+k-1}) = \mathbf{w}$. Now in some cases, some w_is might be in the wrong parts which makes it impossible for any copy of $\mathcal{T}$ to appear with those roots. We will assume we are not in this case, and that the w_is are all in the correct parts so that copies of $\mathcal{T}$ are in fact possible. We will consider these w_is as elements of $\mathbb{F}_q^b$.

Let C be the set of all copies of $\mathcal{T}$ in G that have $w_1, w_2, ..., w_{b-a+k-1}$ as its roots. We are interested in this because whenever we have a copy of a hypergraph of $\mathcal{T}^p$ with the given roots, that implies $|C| \geq p$. For the moment, our goal will be to find an upper bound for $\mathbb{P}(|C| \geq p)$, since that will also be an upper bound on the probability

of getting a copy of $\mathcal{T}^p$.

Lemma 9. *If q is sufficiently large, then $|C| \geq p \Leftrightarrow |C| \geq q/2$.*

Proof: We will treat vertices of our hypergraph as elements in $\mathbb{F}_q^b$. Furthermore, we will identify copies of $\mathcal{T}$ rooted at $\mathbf{w}$ with vectors of the form $(x_1, x_2, ...x_a)$, where the x_is represent the a non-roots in our copy of $\mathcal{T}$ in the correct order.

When is $(x_1, x_2, ..., x_a)$ in C? It is in C if and only if:
(1) all the sets of the form $\{x_j, x_{j+1}, ..., x_{j+k-1}\}$ that correspond to edges of $\mathcal{T}$ are actually edges in G.
(2) all sets of the $\{x_j, x_{j+1}, ..., x_{j+k-2}, w_l\}$ that correspond to edges of $\mathcal{T}$ are actually edges in G.
(3) $x_i \neq x_j$ whenever those two vertices are in the same part
(4) $x_i \neq w_j$ whenever those two vertices are in the same part.

The first condition is equivalent to $\forall i, \forall j,\ f_i(x_j, x_{j+1}, ..., x_{j+k-1}) = 0$ whenever this corresponds to an edge of $\mathcal{T}$. So the set of $\{x_1, x_2, ..., x_a\}$ that satisfy condition (1) form the rational points of a variety made up of at most $a \cdot (a - k + 1)$ equations, each of degree at most d.

The second condition is equivalent to $\forall i, \forall j,\ f_i(x_j, x_{j+1}, ..., x_{j+k-2}, w_l) = 0$ whenever this corresponds to an edge of $\mathcal{T}$. So similarly to the first case, the set of $\{x_1, x_2, ..., x_a\}$ that satisfy condition (2) form the rational points of a variety made up of at most $a \cdot (b - a + k - 1)$ equations, each of degree at most d. Combining conditions (1) and (2) gives a variety V made up of at most $a \cdot b$ equations, each of degree at most d.

The third and fourth condition together make up a system of at most $\binom{a}{2} + a \cdot (b - a + k - 1)$ complements of linear equations, so the set of $(x_1, x_2, ..., x_a)$s that satisfy these conditions is the complement of the rational points of a variety U made up of the product of at most $\binom{a}{2} + a \cdot (b - a + k - 1)$ linear equations.

We have $C \cong V(\mathbb{F}_q) \backslash U(\mathbb{F}_q)$, where V is a variety in a variables defined by at most ab equations of degree at most d, and U is a variety in the same a variables defined by at most $\binom{a}{2} + a \cdot (b - a + k - 1)$ equations of degree 1. We can then split V into a number of irreducible components $V = V_1 \cup V_2 \cup ... \cup V_v$, where $v \leq d^{ab}$. Then $C \cong (V_1(\mathbb{F}_q) \backslash U(\mathbb{F}_q)) \cup (V_2(\mathbb{F}_q) \backslash U(\mathbb{F}_q)) \cup ... \cup (V_v(\mathbb{F}_q) \backslash U(\mathbb{F}_q))$. Now for each irreducible component V_i, either

$V_i \subset U$ (in which case $V_i \backslash U = \emptyset$, so we can ignore this component), or $V_i \cap U$ has dimension strictly smaller than V_i. By the Lang-Weil bound (Theorem 8), $|V_i(\mathbb{F}_q)| = (1 + O(q^{-1/2})) \cdot q^{dim(V_i)}$, while $|V_i(\mathbb{F}_q) \cap U(\mathbb{F}_q)| \leq (1 + O(q^{-1/2})) \cdot q^{dim(V_i \cap U)}$. Therefore when q is large enough, we have $2q^{dim(V_i)} > |V_i(\mathbb{F}_q) \backslash U(\mathbb{F}_q)| \geq \frac{1}{2} q^{dim(V_i)}$. Adding all the pieces up, we have $2v \cdot q^{dim(V)} > |V(\mathbb{F}_q) \backslash U(\mathbb{F}_q)| \geq \frac{1}{2} q^{dim(V)}$. When $\dim(V) \geq 1$, this gives us $|V(\mathbb{F}_q) \backslash U(\mathbb{F}_q)| \geq q/2$. Otherwise, $dim(V) = 0$ and $|V(\mathbb{F}_q) \backslash U(\mathbb{F}_q)| < 2v$.

Since $v \leq d^{ab}$, $2v \leq p$ so $|V(\mathbb{F}_q) \backslash U(\mathbb{F}_q)| < p$ when $dim(V) = 0$. Now the lemma is proved: we either have $|C| \geq q/2$ or $|C| < p$, as required.

□

Continuing on with the main proof, we have $\mathbb{P}(|C| \geq p) = \mathbb{P}(|C| \geq q/2) = \mathbb{P}(|C|^s \geq (q/2)^s)$, which by Markov's inequality is $\leq \frac{\mathbb{E}(|C|^s)}{(q/2)^s}$. We now want to calculate $\mathbb{E}(|C|^s)$. Because $\mathcal{T}^{\leq s}$ was defined to be the set of all graphs you can make by taking the union of s copies of T all rooted at the same place, an element of $|C|^s$ corresponds to a copy of an element H in $\mathcal{T}^{\leq s}$ (obtained by taking the union). Also, for every element H in $\mathcal{T}^{\leq s}$, let $\gamma_s(H)$ be the number of ways of expressing it as a union of s copies of T. This means that:

$$\mathbb{E}(|C|^s)) \leq \sum_{H \in \mathcal{T}^{\leq s}} \gamma_s(H) \cdot \mathbb{E}(|\{A \in \mathrm{Hom}(H, G) \, : \, r(A) = \underline{w}\}|)$$

To get any further, we will need the following lemma:

Lemma 10. *For any $H \in \mathcal{T}^{\leq s}$, we have $\mathbb{E}(|\{A \in \mathrm{Hom}(H, G) \, : \, r(A) = \underline{w}\}|)) = q^{b \cdot (|H| - |R|) - a \cdot e(H)}$. In other words, the expected number of copies of H rooted at $\underline{w}$ is equal to $q^{b \cdot (|H| - |R|) - a \cdot e(H)}$*

Proof: Call $m = |H| - |R|$. We have: $(x_1, x_2, ..., x_m)$ forms a copy of H rooted at $\underline{w}$ if and only if for all i, $f_i(x_{j_1}, x_{j_2}, ..., x_{j_k}) = 0$ whenever this corresponds to an edge of H and for all i, $f_i(x_{j_1}, x_{j_2}, ..., x_{j_{k-1}, w_{j_k}}) = 0$ whenever that corresponds to an edge of H. The f_is are independent from each other so we only need to find, for each i, the probability that $f_i(x_{j_1}, x_{j_2}, ..., x_{j_k}) = 0$ whenever this corresponds to an edge of H and the probability that $f_i(x_{j_1}, x_{j_2}, ..., x_{j_{k-1}, w_{j_k}}) = 0$ whenever that corresponds to an edge of H.

For simplicity, we shall call the $e(H)$ points in $(\mathbb{F}_q^b)^k$ corresponding to edges of H:

$y_1, y_2, ..., y_{e(H)}$ and fix them. We want to calculate $\mathbb{P}(\forall j\ f_i(y_j) = 0)$, knowing that f_i is a random polynomial of degree $\leq d$. We can first without loss of generality make a change of variable π such that the first coordinate of each y_j is different. To do so, we proceed as follows: a change of variable is just a non-singular $bk \times bk$ matrix acting on the y_js. The first coordinates of $\pi(y_j)$ is given by the dot product of y_j with the first row vector of π. Given any j and j', the first coordinate of $\pi(y_j)$ is equal to the first coordinate of $\pi(y_{j'})$ if and only if the elements of first row vector of π satisfy some linear equation. Thus by repeating this operation over all choices of j, j', we get a set of $\binom{e(H)}{2}$ linear equations in bk variables. The set of all possible first rows for π has size $q^{bk} - 1$ (we have bk coordinates and the only thing we require is that not all of them are 0). The set of all possible first rows that satisfy one particular linear equation has size $q^{bk-1} - 1$ (there is some variable that we can express as a function of the $bk-1$ others, and we still disallow the 0). So if we disallow all first rows that satisfy one of the equations, we end up with at least $q^{bk} - 1 - \binom{e(H)}{2}(q^{bk-1} - 1)$ possible first rows of π. Note that because $H \in \mathcal{T}^s$, we have $e(H) \leq sb = d + 1$, and since we assumed that $q > \binom{d+1}{2}$, we have $\binom{e(H)}{2}/q < 1$. Thus, this number is positive, so there is some choice for a first row of π that makes the first coordinate of each $\pi(y_j)$ different. From there, add on the other $bk - 1$ rows of π arbitrarily just making sure that π is invertible. On top of replacing the y_js, we'll also be replacing f_i with $f_i\pi^{-1}$ so that $f_i(y_j)$ stays the same. Note that because f_i was chosen uniformly at random amongst polynomials of degree at most d and because π is a bijection, $f_i\pi^{-1}$'s distribution is also uniform amongst polynomials of degree at most d. Therefore without loss of generality, we can assume that the first coordinate of the y_js are distinct. We'll let z_j be the first coordinate of y_j.

Now suppose we are given a random polynomial of degree at most d : $f(x_1, x_2, ..., x_{kb})$. Consider the coefficients in front of the terms 1 , x_1 , x_1^2 , x_1^3,... and $x_1^{e(H)}$; call them c_0,c_1,...,$c_{e(H)}$ respectively. These c_is are random variables chosen independently and uniformly in $\mathbb{F}_q$. We can write f as:

$$f = c_0 + c_1x_1 + c_2x_1^2 + ... + c_{e(H)}x_1^{e(H)} + f'$$

where f' consists of all the other terms that aren't already written down. By letting $c'_{e(H)-1} = c_{e(H)-1} + c_{e(H)}z_{e(H)}$, we can rewrite $c_{e(H)}x_1^{e(H)} + c_{e(H)-1}x_1^{e(H)-1}$ as $c_{e(H)}x_1^{e(H)-1}(x_1 - z_{e(H)}) + c'_{e(H)-1}x_1^{e(H)-1}$. Note that since $c_{e(H)-1}$ was chosen uniformly at random in $\mathbb{F}_q$ independent of all the other cs and independently of f', $c'_{e(H)-1}$ also has the same properties. We can repeat this process multiple times until we write f as:

$$f = c'_0 + (x_1 - z_1)\left[c'_1 + (x_1 - z_2)\left[c'_2 + (x_1 - z_3)\left[\ldots\left[c'_{e(H)-1} + (x_1 - z_{e(H)})c'_{e(H)}\right]\ldots\right]\right]\right] + f'$$

where all the c'_is are uniformly chosen in $\mathbb{F}_q$ independently of each other and independently of f'.

Suppose we fix f'. The polynomial is 0 at y_1 if and only if $c'_0 = -f'(y_1)$, which has probability $1/q$. Then given that $f(y_1) = 0$, the polynomial is 0 at y_2 if and only if $c'_1 = \frac{c'_0 + f'(y_2)}{z_1 - z_2}$, which also has probability $1/q$ (remember that all the z_is were distinct so we are not dividing by 0). We continue in this fashion by induction until we reach $f(y_{e(H)})$ is 0 with probability $1/q$ given that all the others are also 0. Multiplying everything together, we get that the probability that $f(y_j) = 0$ for all j is $q^{-e(H)}$.

Going back to the last inequality, we get the probability that $(x_1, x_2, ..., x_m)$ forms a copy of H rooted at $\underline{w}$ is equal to $\prod_{i=1}^{a} q^{-e(H)} = q^{-a \cdot e(H)}$. Therefore, the expected number of copies of H rooted at $\underline{w}$ is equal to $q^{b \cdot m - a \cdot e(H)} = q^{b \cdot (|H| - |R|) - a \cdot e(H)}$ and the lemma is proved.

□

Remember from Lemma 8 that for all Hs in $\mathcal{T}^s$, we have $e(H) \geq (|H| - |R|)\frac{b}{a}$, so by combining Lemmas 8 and 10 we get: $\mathbb{E}(|\{A \in \mathrm{Hom}(H, G) \,:\, r(A) = \underline{w}\}|) \leq 1$.

Putting this back in the previous inequality, we have:

$$\begin{aligned} \mathbb{E}(|C|^s) &\leq \sum_{H \in \mathcal{T}^{\leq s}} \gamma_s(H) \cdot \mathbb{E}(|\{A \in \mathrm{Hom}(H, G) \,:\, r(A) = \underline{w}\}|) \\ &\leq \sum_{H \in \mathcal{T}^{\leq s}} \gamma_s(H) \end{aligned}$$

which is a constant depending only on s. We will call this β_s.

Again putting this back into the first inequality, we get: $\mathbb{P}(|C| \geq p) \leq \frac{\mathbb{E}(|C|^s)}{(q/2)^s} \leq \frac{2^s \beta_s}{q^s}$.

At this point, we know that when we pick $w_1, w_2, ..., w_{b-a+k-1}$ at random (in the correct parts), we have a probability of less than $\frac{2^s \beta_s}{q^s}$ of finding a hypergraph of $\mathcal{T}^p$ rooted at $\mathbf{w}$. Let D be the number of choices for $\mathbf{w}$ that do lead to finding such a

hypergraph. $\mathbb{E}(D) \leq k! \cdot (q^b)^{(b-a+k-1)} \cdot \frac{2^s \beta_s}{q^s}$. But now remember that s was defined as $b(b-a+k-1)+a+1$, so we get $\mathbb{E}(D) \leq \frac{k! 2^s \beta_s}{q^{a+1}}$.

At this point we're finally ready to reconsider the cases where G is empty. We can split the expectation of D into the case where G is empty and the case where it is not:
$\mathbb{E}(D) = \mathbb{E}(D \mid G \text{ empty}) \cdot \mathbb{P}(G \text{ empty}) + \mathbb{E}(D \mid G \text{ non-empty}) \cdot \mathbb{P}(G \text{ non-empty})$.
We clearly have no copies of the forbidden hypergraphs when G is empty, so $\mathbb{E}(D \mid G \text{ empty}) = 0$. Meanwhile, we know from earlier that $\mathbb{P}(G \text{ non-empty}) \geq q^{-a}$. Putting this together, we get:
$\mathbb{E}(D \mid G \text{ non-empty}) \leq \mathbb{E}(D) \cdot q^a \leq \frac{k! 2^s \beta_s}{q}$

Now this has order $\Theta(1/q)$ so when q is large enough, we get $\mathbb{E}(D \mid G \text{ non-empty}) < 1$. This proves that there is some choice of $f_1, f_2, ..., f_{b-a+k-1}$ for which G is non-empty but that gives no elements of $\mathcal{T}^p$ inside G.

Thus, we have constructed a hypergraph G with $\Theta(N^{k-\frac{a}{b}})$ edges and that does not contain any element of $\mathcal{T}^p$. Thus, $ex(n, \mathcal{T}^p) = \Omega(n^{k-\frac{a}{b}})$ and the proof of the lower bound is complete.

4.4 The upper bound

Suppose we are given a real number r and a k-hypergraph G with n vertices and $n^{k-r}/k!$ edges. To get the upper bound, we need to find a copy of a graph in $\mathcal{T}^p$ whenever $r \leq O(a/b)$. In [8], they used the fact that given graph, we can pick a subgraph with high minimal codegree. Here, we will notice that $\mathcal{T}$ is a tight k-hypertree (in fact, we used $\mathcal{T}$ as our example of a tight k-hypertree in chapter 2.4) and therefore we can simply apply Corollary 1 of Theorem 2.

Let $c = \sqrt[b]{2[(p-1)a]!/[(p-2)a]!}$ and let X be a k-hypergraph with n vertices and at least $\frac{c}{k} n^{1-r} \binom{n}{k-1}$ edges. The average degree of a $(k-1)$-set in X is $k \cdot \frac{c}{k} n^{1-r} \cdot \binom{n}{k-1} / \binom{n}{k-1} = cn^{1-r}$. $\mathcal{T}$ is a tight k-hypertree with b edges so by applying Corollary 1 of Theorem 2 to this tells us that the number of copies of $\mathcal{T}$ in X is at least:

$$\binom{n}{k-1}(k-1)! \cdot (cn^{1-r})^b \cdot \left(1 - O\left(\frac{\ln(n)}{cn^{1-r}}\right)\right)$$

so as long as n is large enough, this is strictly more than:

$$n^{b-a+(k-1)} \cdot \frac{c^b}{2}$$

Finishing up the upper bound

We know that there are strictly more than $\left(n^{b+k-1-rb}\right)\frac{c^b}{2}$ copies of $\mathcal{T}$ in any large enough hypergraph X with n vertices and $\frac{c}{k}n^{1-r}\binom{n}{k-1}$ edges. Now note that there are only $n^{b-a+k-1}$ possibilities for choosing distinct roots. Therefore, given a random ordered set of $b-a+k-1$ vertices of X, the expected number of copies of $\mathcal{T}$ rooted at them is strictly more than $\frac{c^b}{2}$. This is an average so we can pick a set of $b-a+k-1$ vertices that have above average number of copies of $\mathcal{T}$ rooted at them. Now consider the union of all these $> \frac{c^b}{2}$ copies of $\mathcal{T}$ rooted at the same place. We claim that this forms an element of $\mathcal{T}^u$ for some $u \geq p$.

An element of $\mathcal{T}^{\leq p-1}$ can have at most $(p-1)a$ non-roots, which means that we can find at most $[(p-1)a]!/[(p-2)a]!$ copies of $\mathcal{T}$ in it (just choose the order of the vertices). But now $\frac{c^b}{2} = [(p-1)a]!/[(p-2)a]!$, which means we have too many copies of $\mathcal{T}$ for them to fit into $\mathcal{T}^{\leq p-1}$. Therefore it has to be an element of $\mathcal{T}^{\leq u} \backslash \mathcal{T}^{\leq p-1}$ for some large u.

But now we can remove edges and vertices from our element of $\mathcal{T}^{\leq u} \backslash \mathcal{T}^{\leq p-1}$ until we find an element of $\mathcal{T}^{\leq p} \backslash \mathcal{T}^{\leq p-1} = \mathcal{T}^p$. Therefore we have found an element of $\mathcal{T}^p$ inside X.

Therefore, if there are strictly more than $\frac{\sqrt[b]{2[(p-1)a]!/[(p-2)a]!}}{k} n^{1-r}\binom{n}{k-1} = \Theta(n^{k-r})$ edges in the hypergraph X and n is sufficiently large, then we have a copy of a hypergraph of $\mathcal{T}^p$ inside X. Therefore $ex(n, \mathcal{T}^p) \leq O(n^{k-\frac{a}{b}})$. Combining this with the lower bound we proved in the first section we get $ex(n, \mathcal{T}^p) = \Theta(n^{k-\frac{a}{b}})$ and the proof of Theorem 6 is complete.

□

4.5 The case where $r \geq 1$

We will now try to prove Theorem 7, which is the generalisation of Theorem 6 from $0 \leq r < 1$ to $0 \leq r < k-1$. To do so, we use the following lemma:

Lemma 11. *Given a set of l-hypergraphs $\mathcal{F}$ (each of which contains 2 disjoint edges) and some $k > l$, there exists some set $\mathcal{F}'$ of k-hypergraphs with $ex(n+k-l, \mathcal{F}') = ex(n, \mathcal{F})$ for all n.*

Proof of Lemma 11: For a l-hypergraph F and vertices $x_1, x_2, ..., x_{k-l}$, define the k-hypergraph $(F, x_1, x_2, ..., x_{k-l})$ to have vertices $V(F) \cup \{x_1, x_2, ..., x_{k-l}\}$ and edges $\{E \cup \{x_1, x_2, ..., x_{k-l}\} : E \text{ an edge of } F\}$. We define $(\mathcal{F}, x_1, x_2, ..., x_{k-l}) = \{(F, x_1, x_2, ..., x_{k-l}) : F \in \mathcal{F}\} \cup \{k\text{-hypergraphs with} \leq l+2 \text{ edges that are not of the form } (H, x_1, x_2, ..., x_{k-l}) \text{ for any } H\}$. We claim that $\mathcal{F}' = (\mathcal{F}, x_1, x_2, ..., x_{k-l}))$ will solve the problem.

Suppose G is a l-hypergraph with $ex(n, \mathcal{F})$ edges that doesn't contain any element of $\mathcal{F}$. Consider $(G, y_1, y_2, ..., y_{k-l})$ for some vertices $y_1, y_2, ..., y_{k-l}$. Then first of all, every set of $\leq l+2$ edges of $(G, y_1, y_2, ..., y_{k-l})$ is of the form $(H, y_1, y_2, ..., y_{k-l})$ because every edge contains $y_1, y_2, ..., y_{k-l}$. Now suppose it contains a different element of $(\mathcal{F}, x_1, x_2, ..., x_{k-l})$, say $(F, x_1, x_2, ..., x_{k-l})$. Now F contains 2 edges that do not intersect, so $(F, x_1, x_2, ..., x_{k-l})$ contains two edges that intersect only at $x_1, x_2, ..., x_{k-l}$. Since any two edges of $(G, y_1, y_2, ..., y_{k-l})$ intersect at $y_1, y_2, ..., y_{k-l}$, that must mean that $y_1, y_2, ..., y_{k-l}$ are the images of $x_1, x_2, ..., x_{k-l}$ in some order. Now taking that copy of $(F, x_1, x_2, ..., x_{k-l})$ in $(G, y_1, y_2, ..., y_{k-l})$ and removing $y_1, y_2, ..., y_{k-l}$ from all the edges, we end up with a copy of F inside G. This contradicts our original assumption about G. Therefore $(G, y_1, y_2, ..., y_{k-l})$ is a k-hypergraph with $ex(n, \mathcal{F})$ edges that does not contain any element of $(\mathcal{F}, x)$, so $ex(n+k-l, (\mathcal{F}, x_1, x_2, ..., x_{k-l})) \geq ex(n, \mathcal{F})$, as required.

Conversely, suppose that G is a $k-$*hypergraph* that doesn't contain any element of $(\mathcal{F}, x_1, x_2, ..., x_{k-l})$. Since every set of $l+2$ edges is of the form $(H, x_1, x_2, ..., x_{k-l})$, that means that the entire hypergraph is of the form $(K, x_1, x_2, ..., x_{k-l})$ for some K.

Indeed, suppose for a contradiction that the hypergraph is not of that form. Pick two edges e_1 and e_2. They intersect in at most $k-1$ places. Pick $k-l$ of those and call the set S. Now the hypergraph is not of the form $(H, x_1, x_2, ..., x_{k-l})$, so that means that there is some other edge, e_3, that doesn't contain S. Now e_1,e_2,e_3 intersect in at most $k-2$ places. Repeat the argument several times until we get edges e_1, e_2, e_3, ..., e_{l+2} that intersect in at most $k-l-1$ places. Thus, we have $l+2$ edges that are not of

the form $(H, x_1, x_2, ..., x_{k-l})$, contradicting our assumption. Therefore G must be of the form $(H, x_1, x_2, ..., x_{k-l})$.
Then this K cannot contain any element F of $\mathcal{F}$ because otherwise $G = (K, x_1, x_2, ..., x_{k-l})$ would contain $(F, x_1, x_2, ..., x_{k-l})$. Therefore K has at most $ex(n, \mathcal{F})$ edges, so G also has at most $ex(n, \mathcal{F})$ edges. Thus $ex(n + k - l, (\mathcal{F}, x_1, x_2, ..., x_{k-l})) \leq ex(n, \mathcal{F})$

This proves that $ex(n, \mathcal{F}) = ex(n + k - l, (\mathcal{F}, x_1, x_2, ..., x_{k-l}))$, completing the proof of the lemma.

□

Proof of Theorem 7: Given a rational $r, 0 \leq r < k-1$, let $l = \lceil k-r \rceil$, and $r' = l-k+r$, so $0 \leq r' < 1$. By Theorem 6, we know that there exists a set of l-hypergraphs $\mathcal{F}$ with $ex(n, \mathcal{F}) = \Theta(n^{l-r'})$. Remember in the proof of Theorem 6 that if $b \geq k$ (which we could assume without loss of generality), then there were at least 2 disjoint edges in every hypergraph of $\mathcal{F}$. Now by applying Lemma 11, we get some set of k-hypergraphs $\mathcal{F}'$ with $ex(n, \mathcal{F}') = \Theta(n^{l-r'}) = \Theta(n^{k-r})$.

□

Remarks:

The case $k > r > k-1$ is impossible :
Suppose that $\mathcal{F}$ is a collection of k-graphs which has $ex(n, \mathcal{F}) = \Theta(n^{k-r})$ for some $k > r > k-1$.

Now consider X to be the k-hypergraph with n vertices defined as follows: it consists of some set S of t vertices, for some $0 \leq t \leq k-1$. The other $n-t$ vertices are partitioned into $\lfloor (n-t)/(k-t) \rfloor$ sets of size $k-t$, which we will call $e_1, e_2, ..., e_{\lfloor (n-t)/(k-t) \rfloor}$. The edges of the hypergraph are exactly $e_i \cup S$ for $1 \leq i \leq \lfloor (n-t)/(k-t) \rfloor$. This hypergraph X has the property that the intersection of any two edges is exactly S therefore it is a sunflower. It also has $\Theta(n)$ edges, which is larger than $c \cdot n^{k-r}$ for large enough n. Therefore $\mathcal{F}$ must contain a subgraph of X. However, any subgraph of X must also have the property that any the intersection of two edges is exactly S, i.e. it is another

sunflower. We will call this sunflower F_t.

In this way, we get for all $0 \leq t \leq k-1$, a sunflower F_t in $\mathcal{F}$ with kernel size t. The Sunflower Lemma [11] states that when this occurs, $ex(n, \mathcal{F})$ has order $O(1)$. This contradicts our assumption that $ex(n, \mathcal{F}) = \Theta(n^{k-r})$. Therefore it is indeed impossible to have a collection of k-hypergraphs with $ex(n, \mathcal{F}) = \Theta(n^{k-r})$ for any $k > r > k-1$.

The case $r = k$ is possible: The hypergraph family consisting of every sunflower with 2 edges has $ex(n, E) = 1$.

The case $r = k-1$ is possible: Firstly, when $k = 2$, then we claim that the path with 3 edges, P_3, has $ex(n, P_3) = n$ or $n-1$.

Indeed, suppose we are given a connected component in a P_3-free graph G. This connected component is either just a single vertex x or an edge xy or it contains 2 incident edges, say xy and yz. If there is another edge in the component incident to x or z, then it has to be incident to both because otherwise we have a P_3, so we get a triangle $\{xy, yz, zx\}$ and we can no more edges to this. If we are not in the triangle case, then all additional edges have to be incident to y. So we get a star centred at y: $\{yz, yz, yt_1, yt_2, ...\}$.

So every component of G is either a triangle (3 vertices and 3 edges) or a star ($k+1$ vertices and k edges for some $k \geq 0$). So the number of edges has to be less than the number of vertices. If the number of vertices n is divisible by 3, then $ex(n, P_3) = n$ because we can use only triangles. Otherwise, we have to use a star at some point which gives us $ex(n, P_3) = n - 1$.

For larger k, we simply apply Lemma 11 to get a collection of k-hypergraphs $\mathcal{F}$ with $ex(n, \mathcal{F}) = \Theta(n)$ as required.

So in conclusion, the rationals r for which there exist some finite $\mathcal{F}$ with $ex(n, \mathcal{F}) = \Theta(n^{k-r})$ are exactly those in the set: $\{r \in \mathbb{Q} : 0 \leq r \leq k-1\} \cup \{k\}$.

Chapter 5

Implicit representation conjecture for semi-algebraic graphs

5.1 Introduction

Definition 23. *Suppose we are given a Euclidean space $\mathcal{S}$, a finite set of symmetric polynomials $f_1, f_2, ..., f_k$ on $\mathcal{S} \times \mathcal{S}$, and a sentence $\mathcal{T}$ whose atomic formulae are $f_1 \geq 0, f_2 \geq 0, ..., f_k \geq 0$. The semi-algebraic family of graphs associated with $\mathcal{T}$ is the set of all finite graphs whose vertices are points $s \in \mathcal{S}$ and whose edges are exactly those pairs of points s, s' that satisfy $\mathcal{T}(s, s')$.*

Notice that $f \leq 0$ is equivalent to $-f \geq 0$, that $f = 0$ is equivalent ($f \geq 0$ and $-f \geq 0$), that $f > 0$ is equivalent to ($f \geq 0$ and not $-f \geq 0$) and similarly for $f < 0$. Therefore these formulas are also allowed.

For an example of a semi-algebraic family, the family of disk graphs consists of all graphs whose vertices are closed disks in the plane and where edges indicate that two disks intersect. The vertices can be viewed as points in $\mathbb{R}^3$: (x, y, r) where (x, y) are the coordinates of the center of the disk and r is the radius. There is an edge between (x_1, y_1, r_1) and (x_2, y_2, r_2) if and only if $(x_1 - x_2)^2 + (y_1 - y_2)^2 \leq (r_1 + r_2)^2$.

The family of unit disk graphs is defined in a similar way except all the radii are 1.

Essentially, semi-algebraic families of graphs are made up of graphs that are defined geometrically or algebraically. They are very useful in graph theory because they are a good way of constructing graphs with certain properties. Work on semi-algebraic graphs

has mostly been focused on specific families, such as the aforementioned family of disk graphs, which has applications in computational geometry [7]. However, there are a few general results. In 2005, Alon, Pach, Pinchasi, Radoičić and Sharir proved that given any semi-algebraic family of graphs, that every graph in it with n vertices contains two subsets of vertices of size ϵn (where ϵ is a constant), such that either all edges between them or no edges between them. They also proved that there exists either a complete subgraph of size n^δ or an induced empty subgraph of size n^δ (where δ is another constant). [2]

In 2013, Blagojević, Bukh and Karasev looked at algebraic methods while trying to solve the Turán problem for the complete bipartite graph $K_{s,s}$ and showed that one particular 'natural' type of semi-algebraic graph cannot be used to construct a $K_{s,s}$-free graph with $\Theta(n^{2-1/s})$ edges [4].

The problem we are trying to solve in this chapter is a special case of the Implicit Representation conjecture, first posed by Kannan, Naor and Rudich in 1992 [16], which was also asked by Spinrad in 2003 [27]. We want to come up with a method for storing graphs using the least number of bits per vertex. A hereditary family of graphs is one in which all induced subgraphs of every graph in the family are also in the family. Given such a hereditary family of graphs $\mathcal{G}$ (for example: disk graphs or unit disk graphs), let $\mathcal{G}^{(n)}$ mean the subset of graphs which have exactly n vertices. For every n, we want a function $F^{(n)} : \mathcal{G}^{(n)} \to [2^m]^n$, and a symmetric function $G^{(n)} : [2^m] \times [2^m] \to \{0, 1\}$ such that for every graph $H \in \mathcal{G}^{(n)}$ and every pair of vertices i, j in H, we have $G(F(H)_i, F(H)_j) = 1$ if and only if there is an edge between i and j. Furthermore, we want to minimise $m = m(n)$, which is the amount of information per vertex. The Implicit Representation Conjecture states that if there exists a constant c such that the family $\mathcal{G}$ contains less than $2^{cn \ln(n)}$ graphs of size n for all n, then there exists a constant c' such that $m = c' \log_2(n)$ will be sufficient for every graph of size n in the family. The Implicit Representation Conjecture has been proved for a large number of families by Atminas, Collins, Lozin and Zamaraev in [3].

A corollary of Warren's Theorem (1968) [30, 1] shows that semi-algebraic families do indeed have at most $2^{O(n \ln(n))}$ graphs of size n, so do satisfy the conditions for the Implicit Representation Conjecture. We'll see the derivation of this corollary at the end of section 5.2.

The trivial lower bound for this problem matches the conjecture, at $m = \log_2(n)$ since that is the amount of information required to identify a vertex amongst n. More specifically, if we use less than $\log_2(n)$ bits, then there are less than n possible options for what the data can be, so by the pigeon-hole principle, there exist two vertices i and j with $F(H)_i = F(H)_j$. This means that their neighbourhoods are identical. However, if we let the graph be a path, then every vertex has a different neighbourhood, which is a contradiction. If the family is defined by the intersection of bounded non-trivial shapes (such as the disk graph), then it is fairly easy to see that we can draw a path using these shapes.

A trivial upper bound that works for all graphs is $m = \lceil \frac{n-1}{2} \rceil + \lceil \log_2(n) \rceil$. To achieve this, we write the vertices as $0, 1, ..., n-1$ in $\mathbb{Z}/n\mathbb{Z}$, and store this information using $\lceil \log_2(n) \rceil$ bits. Then for every vertex i, let $F(H)_i$ be a list of $\lceil \frac{n-1}{2} \rceil$ 0s and 1s, with a 1 in the kth position if and only if there is an edge between i and $i+k$. For every pair i and j, G will then output the $(j-i)$th entry of $F(H)_i$ if $j-i$ is between 1 and $\lceil \frac{n-1}{2} \rceil$ and otherwise it will output the $(i-j)$th entry of $F(H)_j$.

A natural idea we could have would be to store integer approximations of the coordinates of all the vertices. This looks like a good idea because it is easy to store integers, and because the function $G^{(n)}$ is easy to compute (just evaluate all the polynomial inequalities). Unfortunately this doesn't work. In 2011, McDiarmid and Muller [22] proved that there exist unit disk graphs with n vertices but for which every realisation of it on the plane had to have four vertices a, b, c, d for which $\frac{|a-b|}{|c-d|} > 2^{2^{\Omega(n)}}$. If a, b, c, d had integer coordinates, then one of a or b has to have a coordinate of size at least $2^{2^{\Omega(n)}}$. This requires $2^{\Omega(n)}$ bits to store which is even more than the trivial bound.

In 2012, Kang and Muller [15] improved upon this result in two ways. Firstly by showing that the dimension k of the ambient space can be arbitrary, and secondly by replacing the integer approximations by rational approximations. They showed that for any $k \geq 2$, there exist unit k-ball graphs with n vertices but for which every realisation of it in $\mathbb{R}^k$ had to have four vertices a, b, c, d for which $\frac{|a-b|}{|c-d|} > 2^{2^{\Omega(n)}}$. (A k-ball graph is defined the same as a disk graph except that the ambient space is of dimension k instead of 2.) Then if a, b, c, d had rational coordinates, then one of these four points has to have a coordinate with numerators or denominators of size at least $\sqrt[4]{2^{2^{\Omega(n)}}} = 2^{2^{\Omega(n)}}$. This requires $2^{\Omega(n)}$ bits to store, which is even more than the trivial upper bound. Thus, storing rational approximations of the coordinates of all the vertices doesn't work in general.

In the first part, we will go even further, and ask whether we can store the coordinates as algebraic numbers instead of rational numbers or integers. However, this runs into the same problems, as we'll see shortly.

In our second part, we find a very minor improvement on the upper bound that does work. It uses a result by Yao and Yao [31], and ideas about semi-algebraic sets from Alon, Pach, Pinchasi and Radoicic [2].

Theorem 9. *Given a semi-algebraic family of graphs $\mathcal{G}$, there exists some constant $\epsilon > 0$ such that we can store every graph of size n in the family using $n^{1-\epsilon}$ bits per vertex for n sufficiently large.*

More specifically, for $m = n^{1-\epsilon}$ there exists a series of functions $F^{(n)} : \mathcal{G}^{(n)} \to [2^m]^n$, and a symmetric function $G^{(n)} : [2^m] \times [2^m] \to \{0, 1\}$ such that for every graph $H \in \mathcal{G}^{(n)}$ and every pair of vertices i, j in H, we have $G(F(H)_i, F(H)_j) = 1$ if and only if there is an edge between i and j.

5.2 Semi-algebraic graphs

5.2.1 Simplification of the problem

Secondly, we will note that we can reduce to the case where the semi-algebraic family is defined by only one inequality. Indeed, if we have a semi-algebraic family defined by $k > 1$ inequalities $f_1(x, y) \geq 0, f_2(x, y) \geq 0, ..., f_k(x, y) \geq 0$ and we have a graph G in this family with vertices $x_1, ..., x_n \in \mathbb{R}^q$. Then the edge set of G can be viewed as the intersection of the edge sets of k semi-algebraic graphs $G_1, G_2, ..., G_k$, each with the same vertex set and defined by inequalities $f_1(x, y) \geq 0$, $f_2(x, y) \geq 0$, ... , $f_k(x, y) \geq 0$ respectively. If we can store each of these G_is using m bits per vertex, then by concatenation, we can store G using $k \cdot m$ bits per vertex. So without loss of generality, we can assume that there is only a single polynomial inequality $f(x, y) \geq 0$ that defines the semi-algebraic family.

Also note that the complement of a graph can be stored using the same number of bits as the original graph, by simply exchanging 0 and 1 in the output of the function G. Therefore we can without loss of generality assume the single polynomial inequality is of the form $f(x, y) \geq 0$.

Now suppose we have a semi-algebraic family of graphs whose vertices can be written as living in the space $\mathbb{R}^q$ and where for any $x, y \in \mathbb{R}^q$, (x, y) is an edge if and only if $f(x, y) \geq 0$ (where f is a polynomial). Let d be the degree of f. Now for every vertex $x = (x_1, x_2, ..., x_q)$ in $\mathbb{R}^q$, we can replace it with a point $\tilde{x}$ consisting of all the terms of degree less than or equal to d, i.e.: $\tilde{x} = (1, x_1, x_2, ..., x_q, x_1^2, x_1x_2, x_1x_3, ..., x_q^2, x_1^3, x_1^2x_2, ..., x_q^3, ..., x_q^d)$. This point exists in the space $\mathbb{R}^{\binom{q+d}{d}}$. We can then also rewrite the polynomial $f(x, y)$ as a bilinear function of $\tilde{x}$ and $\tilde{y}$: $f(x, y) = \tilde{x}^T M \tilde{y}$ where M is a matrix (M is symmetric because f was symmetric). Let Q be the dimension of M.

Note that given a fixed $\tilde{y}$ with $M\tilde{y} \neq 0$, the set of solutions to the equation $z^T M\tilde{y} \geq 0$ forms a half-space. So whenever $M\tilde{y} \neq 0$, the vertices adjacent to y are exactly those in this half-space. If $M\tilde{y} = 0$, then y is adjacent to every vertex of G. We can think of these as living in some half-space whose boundary is "far away". So regardless of which case we're in, given any vertex y, the vertices adjacent to y are exactly those in some half-space.

At this point, we already have everything we need about semi-algebraic graphs to complete the proof; however, the ϵ that we will get in the final result will be a function of the dimension Q of M so decreasing Q will improve the result slightly. So there is one more thing we can do: it is a standard property of bilinear forms that we can diagonalise them, and furthermore, we can make it such that there are only 1s, -1s and 0s on the diagonal. So without loss of generality, we can assume that

$$M = \begin{pmatrix} 1 & 0 & 0 & ... & 0 & 0 & ... & 0 \\ 0 & 1 & 0 & ... & 0 & 0 & ... & 0 \\ 0 & 0 & 1 & ... & 0 & 0 & ... & 0 \\ ... & ... & ... & ... & ... & ... & ... & \\ 0 & 0 & 0 & ... & -1 & 0 & ... & 0 \\ 0 & 0 & 0 & ... & 0 & -1 & ... & 0 \\ ... & ... & ... & ... & ... & ... & ... & ... \\ 0 & 0 & 0 & ... & 0 & 0 & ... & 0 \end{pmatrix}$$

We can delete those coordinates for which M has a zero on the diagonal because they do not impact the result. So without loss of generality, M is a diagonal matrix with only 1s and -1s on the diagonal. In particular, there are only $Q + 1$ types of matrix in

dimension Q. So solving the problem for just these few special cases is enough. Every other semi-algebraic family is a combination of matrices of this type after a change of basis.

Example 1:

For the unit disk graph in the plane, every vertex can be identified with its center: (x, y). Then two disks (x_1, y_1) and (x_2, y_2) intersect if and only if $(x_1-x_2)^2+(y_1-y_2)^2 \leq 4$. There is only a single inequality, so if we put this in bilinear form, we get a single matrix:

$$(x_1^2, x_1y_1, y_1^2, x_1, y_1, 1) \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & -1 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & -1 \\ 0 & 0 & 0 & 2 & 0 & 0 \\ 0 & 0 & 0 & 0 & 2 & 0 \\ -1 & 0 & -1 & 0 & 0 & 4 \end{pmatrix} \begin{pmatrix} x_2^2 \\ x_2y_2 \\ y_2^2 \\ x_2 \\ y_2 \\ 1 \end{pmatrix} \geq 0$$

We can use a change of basis and then delete irrelevant coordinates to turn this matrix into:

$$\mathbf{x}'^T \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & -1 \end{pmatrix} \mathbf{y}' \geq 0$$

Note that the dimension of this matrix is $Q = 4$.

Example 2:

For the disk graph in the plane, every vertex can be identified with its center and its radius: (x, y, r). Then two disks (x_1, y_1, r_1) and (x_2, y_2, r_2) intersect if and only if $(x_1 - x_2)^2 + (y_1 - y_2)^2 \leq (r_1 + r_2)^2$. If we put this in bilinear form, we get

$$(x_1^2, y_1^2, r_1^2, x_1, y_1, r_1, 1) \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 & -1 \\ 0 & 0 & 0 & 0 & 0 & 0 & -1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 2 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 2 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 2 & 0 \\ -1 & -1 & 1 & 0 & 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} x_2^2 \\ y_2^2 \\ r_2^2 \\ x_2 \\ y_2 \\ r_2 \\ 1 \end{pmatrix} \geq 0$$

Using the change of basis, we can replace this by:

$$\mathbf{x}'^T \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & -1 \end{pmatrix} \mathbf{y}' \geq 0$$

So in this case we have $Q = 5$.

5.2.2 Proof that semi-algebraic families satisfy the conditions for the Implicit Representation Conjecture

For this, we use Warren's Theorem [30] [see for example [1, p.1763]]:

Theorem 10 (Warren, 1968)**.** *Suppose we have a set of k real polynomials in l variables of degree at most d and $k \geq l$. If we split $\mathbb{R}^l$ into regions depending on the signs of all the polynomials (i.e.: whether each polynomial is negative, positive or 0 at a given point in $\mathbb{R}^l$), we end up with at most $(8\mathbf{e}dk/l)^l$ regions (where $\mathbf{e}$ is Euler's constant).*

Suppose we have a semi-algebraic family $\mathcal{G}$, with associated Euclidean space $\mathcal{S}$ and associated set of polynomial inequalities $\mathcal{P}$ and also let n be an integer. We want to count the number of graphs in our family with n vertices. A graph G is in $\mathcal{G}$ if and only if there exist n distinct points $x_1, x_2, ..., x_n$ in $\mathcal{S}$, such that $\mathcal{P}(x_i, x_j)$ is true if and only if (x_i, x_j) is an edge of G. So let $\mathcal{Q}_G$ be the set of polynomial inequalities: $\bigcup_{(i,j) \text{ edge}} \mathcal{P}(x_i, x_j) \cup \bigcup_{(i,j) \text{ non-edge}} \neg\mathcal{P}(x_i, x_j)$. Then G is in $\mathcal{G}$ if and only if there exist $x_1, x_2, ..., x_n$ in $\mathcal{S}$ that satisfy $\mathcal{Q}_G(x_1, x_2, ..., x_n)$. Notably, we can split $\mathcal{S}^n$ into regions depending on the signs of the polynomials of $\mathcal{Q}_G$, and then each region will have a unique

graph associated with it.

So how many regions are there? $\mathcal{Q}_G$ is a set of $\binom{n}{2}|\mathcal{P}|$ polynomial inequalities. The number of variables of these polynomials is $n \cdot \dim(\mathcal{S})$, and the maximum degree is d. So by Warren's Theorem, the number of regions is at most $\left(\frac{8\mathbf{e}d\binom{n}{2}|\mathcal{P}|}{n\cdot\dim(\mathcal{S})}\right)^{n\cdot\dim(\mathcal{S})}$ as long as n is large enough. This is less than $(4\mathbf{e}d|\mathcal{P}|n/\dim(\mathcal{S}))^{n\cdot\dim(\mathcal{S})} \leq 2^{cn\ln(n)}$ for some large enough constant c. This completes the proof and shows that semi-algebraic families do in fact satisfy the hypothesis of the Implicit Representation Conjecture.

5.3 The 'algebraic points' method doesn't work for disk graphs

A natural thing we can try to store disk graphs is to let the centers and radii of all the circles be algebraic and just store these numbers. However, this turns out to be worse than the trivial bound. This builds upon the paper [22] where they prove that storing the centres and radii of all the circles as rational numbers doesn't work.

An important part of the proof is that there exists an infinite family of disk graphs such that for any disk representation of them, there are 4 centers x,y,z and t such that $\frac{|x-y|}{|z-t|} > 2^{2^{\Omega(n)}}$. This family was constructed in [22]. We claim that such an object requires at least $\Omega(n)$ bits if the centres are algebraic.

First, how does one store an algebraic number? If x is an algebraic number, it has a minimal integer polynomial it is a solution to: $\sum_{i=0}^{i_{max}} a_i x^i = 0$. We can store each a_i in $\Theta(\log_2(|a_i|+1))$ bits. Therefore storing the polynomial takes $\Theta(\sum_{i=0}^{k}\log_2(|a_i|+1)) + \Theta(i_{max})$ bits. The polynomial also has i_{max} solutions so we additionally need $\Theta(\log_2(i_{max}))$ bits to indicate which solution it is. Therefore it overall takes $\Theta(\sum_{i=0}^{i_{max}}\log_2(|a_i|+1)) + \Theta(i_{max})$ bits to store an algebraic number x. We'll call this number $m(x)$.

For simplicity, we'll consider the real plane on which our disk graph is drawn to be $\mathbb{C}$, so each center only requires a single algebraic number to describe it.

Pick some integer m. What is the largest we can make $|x-y|$ given that $m(x) \leq m$ and $m(y) \leq m$? First of all, we know that $|x-y| \leq |x| + |y|$. Next, suppose that x satisfies $\sum_{k=0}^{k_{max}} c_k x^k = 0$. Then if $|x| > \sum_{k=0}^{k_{max}-1} |c_k|$, we have $|c_{k_{max}} x^{k_{max}}| > \left(\sum_{k=0}^{k_{max}-1} |c_k|\right) x^{k_{max}-1} \geq \sum_{k=0}^{k_{max}-1} |c_k x^k| \geq |\sum_{k=0}^{k_{max}-1} c_k x^k|$ which contradicts the fact that $\sum_{k=0}^{k_{max}} c_k x^k = 0$. Therefore we must have $x \leq \sum_{k=0}^{k_{max}-1} |c_k|$. This is equal to $2^{\log_2((\sum_{k=0}^{k_{max}-1} |c_k|)} \leq 2^{\sum_{k=0}^{k_{max}-1} \log_2(c_i)} \leq 2^{m-1}$. So the most we can have $|x-y|$ be is 2^m.

Now what is the smallest we can make $|x-y|$ for $x \neq y$? First, we will write $x-y$ as the solution to a polynomial. Say x is a solution to the polynomial $\sum_{i=0}^{i_{max}} a_i x^i$ while y is a solution to the polynomial $\sum_{j=0}^{j_{max}} b_j y^j$. Consider $(a_{i_{max}} b_{j_{max}} (x-y))^l$ for some integer l. We can develop it into $(a_{i_{max}} b_{j_{max}})^l \sum_{k=0}^{l} \binom{l}{k} x^k (-y)^{l-k}$. The sum of the absolute values of the coefficients is $|a_{i_{max}} b_{j_{max}}|^l \cdot 2^l$.

Now what we do is, starting from $k = l$ and going down to $k = i_{max}$, we replace all instances of $a_{i_{max}} x^k$ with $-\sum_{i=0}^{i_{max}-1} a_i x^{k-i_{max}+i}$. Note that because we started with $a_{i_{max}}^l$ in every coefficient, we will be able to do this operation l times, which is bigger than the $l - i_{max} + 1$ required. So this will only halt when the only instances of x have exponent less than i_{max}. What does this do to the sum of the absolute values of the coefficients? Well every time we do this operation, we multiply it by at most $\frac{\sum_{i=0}^{i_{max}-1} |a_i|}{|a_{i_{max}}|}$. We know that $\sum_{i=0}^{i_{max}-1} \log_2(|a_i| + 1) \leq O(m)$ so $\sum_{i=0}^{i_{max}-1} |a_i| \leq 2^{O(m)}$ by concavity of the $\log_2$ function. Since we started with the sum of the absolute values of the coefficients at most $|a_{i_{max}} b_{j_{max}}|^l * 2^l$ and we do this operation $l - i_{max} + 1$ times, we end up with the sum of the absolute values of the coefficients is at most $|b_{j_{max}}|^l * |a_{i_{max}}|^{i_{max}-1} * (2^{O(m)})^{l-i_{max}+1}) * 2^l$. Note also that $|a_{i_{max}}| \leq 2^{O(m)}$ so we end up with the sum of the absolute values of the coefficients is at most $2^l * |b_{j_{max}}|^l \cdot 2^{O(ml)}$

We do the same operation with y, to end up with a linear formula for $(a_{i_{max}} b_{j_{max}})^l (x-y)^l$ in terms of $\{x^i y^j | i < i_{max}\,;\, j < j_{max}\}$, and where the sum of all the absolute values of all the coefficients is at most $2^{O(ml)}$.

Now if we do this for all l between 0 and $i_{max} j_{max}$, then we have $i_{max} j_{max} + 1$ formulae inside the linear space generated by $\{x^i y^j | i < i_{max}\,;\, j < j_{max}\}$. But this space has dimension $i_{max} j_{max}$, so our formulae must be linearly dependent. Remembering that each of our formulae represented some power of $a_{i_{max}} b_{j_{max}} (x-y)$, this linear dependence is equivalent to an integer polynomial of degree $i_{max} j_{max}$ that is zero when

evaluated at $a_{i_{max}}b_{j_{max}}(x-y)$. Without loss of generality suppose that this polynomial is minimal; say it has degree d. We'll write this polynomial as $\mu(a_{i_{max}}b_{j_{max}}(x-y))^d = \sum_{k=0}^{d-1}\lambda_k(a_{i_{max}}b_{j_{max}}(x-y))^k$ where μ and all the λs are integers. How big are the coefficients of this polynomial?

We can work out what they are. Since the polynomial was chosen to be minimal, we know that the formulas for $(a_{i_{max}}b_{j_{max}}(x-y))^k$, $k<d$, are all linearly independent. We can list all these formulas in an $i_{max}j_{max} \times d$ matrix of integers which we'll call M, where the rows are linearly independent:

$$\begin{pmatrix} 1 \\ a_{i_{max}}b_{j_{max}}(x-y) \\ (a_{i_{max}}b_{j_{max}}(x-y))^2 \\ ... \\ (a_{i_{max}}b_{j_{max}}(x-y))^{d-1} \end{pmatrix} =$$

$$\begin{pmatrix} 1 & 0 & 0 & 0 & ... & 0 \\ 0 & a_{i_{max}}b_{j_{max}} & -a_{i_{max}}b_{j_{max}} & 0 & ... & 0 \\ 0 & 0 & 0 & a_{i_{max}}^2 b_{j_{max}}^2 & ... & 0 \\ . & . & . & . & ... & . \\ . & . & . & . & ... & . \end{pmatrix} \begin{pmatrix} 1 \\ x \\ y \\ x^2 \\ xy \\ y^2 \\ ... \\ x^{i_{max}-1}y^{j_{max}-1} \end{pmatrix}$$

Meanwhile, we also have a similar formula for $(a_{i_{max}}b_{j_{max}}(x-y))^d$, which takes the form of a vector of integers of size $i_{max}j_{max}$. We'll call this vector $\mathbf{v}$:

$$(a_{i_{max}}b_{j_{max}}(x-y))^d \quad = \quad \mathbf{v} \, . \begin{pmatrix} 1 \\ x \\ y \\ x^2 \\ xy \\ y^2 \\ ... \\ x^{i_{max}-1}y^{j_{max}-1} \end{pmatrix}$$

This formula is a linear combination of the rows of the above matrix: $\forall l\ ,\ \mu \mathbf{v}_l = \sum_{k=0}^{d-1} \lambda_k M_{k,l}$. If we let the vector of λ_ks be λ (of length d), this formula can be rewritten in vector and matrix form as:

$$\mu \mathbf{v} = \lambda M$$

Now pick some linearly independent subset C of the columns of the matrix of size d. This gives us a $d \times d$ non-singular matrix M' Let π_C be the matrix of the orthogonal projection from the space generated by $\{x^i y^j | i < i_{max}\,;\, j < j_{max}\}$ onto the space generated by C. Thus, $M' = M\pi_C$. Also let **v'** be the image of **v** via this projection, i.e. $\mathbf{v}' = \mathbf{v}\pi_C$. The above linear combination continues to hold after projection:

$$\mu \mathbf{v}' = \lambda M'$$

But now we can find out exactly what our λ_ks are by simply using the equation: $\lambda = \mu M'^{-1}$**v'** (remember that M' is non-singular).

We set $\mu = \det(M')$, which will make all the λ_ks be integers. We know that the sum of the absolute values of all the coefficients of in each row of M and **v** are at most $2^{O(ml)}$, so we get that $det(M')$ is at most $d!(2^{O(ml)})^d = 2^{O(md^2)}$. Moreover, for every $k < d$ each λ_k is the determinant of a minor of M' so is also at most $2^{O(md^2)}$.

Putting this all together (remembering that $d \leq i_{max} j_{max}$), we get an integer polynomial that is 0 at $a_{i_{max}} b_{j_{max}} (x - y)$, that has of degree at most $i_{max} j_{max}$, and where all the coefficients are at most $2^{O(m[i_{max} j_{max}]^2)}$. Now both i_{max} and j_{max} are $\leq O(m)$ so this means the polynomial is of degree at most $O(m^2)$ with coefficients at most $2^{O(m^5)}$. Say this polynomial is $\sum_{k=0}^{k_{max}} c_k (x - y)^k$.

If we assume that $|x - y| < \frac{1}{\sum_{k=1}^{k_{max}} |c_k|}$, then $|\sum_{k=1}^{k_{max}} c_k (x-y)^k| \leq \left[\sum_{k=1}^{k_{max}} |c_k|\right] |x - y| < 1 \leq |c_0|$, which contradicts the polynomial being 0. Therefore $|x - y| \geq \frac{1}{\sum_{k=1}^{k_{max}} |c_k|} = \frac{1}{O(m^2)} 2^{-O(m^5)} = 2^{-O(m^5)}$

The ratio between the smallest possible value of $|x - y|$ and the largest is thus of order $2^{O(m^5)} * 2^{O(m)} = 2^{O(m^5)}$. When we use the special graph whose largest ratio is always at

least $2^{2^{\Omega(n)}}$, we get that m must be of order at least $2^{\Omega(n)}$. This is worse than our trivial upper bound of $m = (\frac{1}{2} + o(1))n$.

5.4 An improvement on the upper bound

5.4.1 The case where $f(x, y) \neq 0$ for all vertices x,y

We'll assume for the moment that for every pair of vertices (x, y), we never have $f(x, y) = 0$; we shall deal with that case at the end.

In this case, for every vertex, the set of vertices adjacent to it is just a half-plane, which has a hyperplane as boundary. This is useful because it means we can use the following theorem:

Theorem 11 (Yao and Yao, 1985 [31])**.** *Given a continuous and everywhere positive probability density function on $\mathbb{R}^Q$, there exists a partition of $\mathbb{R}^Q$ into 2^Q regions, each with mass equal to $1/2^Q$ such that every hyperplane in $\mathbb{R}^Q$ must not intersect the interior of at least one of these regions.*

Moreover, these regions are convex polyhedral cones and all the cones have a common apex, called the center.

A corollary of this theorem is the discrete version of it:

Lemma 12. *Given a finite set V of n points in $\mathbb{R}^Q$, there exists a partition of $\mathbb{R}^Q$ into 2^Q regions, each of which contains between $\lfloor n/2^Q \rfloor$ and $\lfloor n/2^Q \rfloor + 2^Q - 1$ of the points, such that every hyperplane in $\mathbb{R}^Q$ must avoid at least one of the interiors of a region.*

Moreover, these regions are convex polyhedral cones and all the cones have a common apex, called the center.

Proof of the lemma: Pick some small $\epsilon > 0$. For every point $x \in V$, we'll have a continuous density function on the ball of radius ϵ centred at x whose total weight is $(1 - \epsilon)/n$. We'll also have a continuous everywhere positive density function of total weight ϵ. Adding up all of these together gives a continuous everywhere positive probability density function on $\mathbb{R}^Q$, which means we can apply Yao and Yao's Theorem. This splits the space into 2^Q convex polyhedral cones with a common apex, and such that each has total weight $1/2^Q$. Let A_ϵ be the region inside the convex hull of the

collection of balls of radius ϵ (i.e.: A_ϵ is a bounded convex region of weight at least $1-\epsilon$).

Suppose we are given an ϵ, together with a polyhedral decomposition as in the lemma. For every vertex of V, we say it borders a certain region if the ball of radius ϵ centred around x intersects the region. The information about which vertices border which regions will be called the configuration of the polyhedral decomposition. A single vertex has at most 2^{2^Q} possible configurations, so at most $n^{2^{2^Q}}$ possible configurations in total. This is finite therefore as $\epsilon \to 0$, there exists a configuration C that occurs infinitely often. So we can pick a decreasing sequence of ϵs together with a corresponding collection of polyhedral cones in configuration C.

Since every face has to be between two regions, the number of faces is at most $\binom{2^Q}{2}$. So we can also pick some subsequence where all the decompositions have the same number of faces. Then there are only a finite number of ways in which to put these faces together (i.e.: which face goes next to which face) so we can again pick some subsequence where the decompositions all have the same number of faces and are arranged the same way.

The remaining polyhedral decompositions all have centres (reminder: the center of a polyhedral decomposition is the common apex of all its faces) by Yao and Yao's Theorem. These centres will stay inside A_ϵ. To see why, assume not, and pick a tangent hyperplane T to A_ϵ that separates it from the center. Then every region of the polyhedral decomposition has weight at least 2^{-Q} so when $\epsilon < 2^{-Q}$, every region has to contain stuff within A_ϵ. Since every region also has an apex at the center, that means every region has to cross T. That means T is a hyperplane that fails to avoid a region, contradicting Yao and Yao's Theorem. Therefore the center has to be within the bounded region A, so there is a subsequence of ϵs such that the centres converge to some point M.

Now a given face of a decomposition in our sequence is a part of a hyperplane that passes through the center of the decomposition. Moreover these centres converge to M so the hyperplanes eventually have to pass within some small distance $\delta > 0$ of M. Since the space of hyperplanes passing within δ of M is compact, there is a hyperplane H passing through C and a subsequence of decompositions such that our given face converges to a part of H. Repeat for all the other faces of the decomposition.

If we take M together with all the hyperplanes passing through it that we constructed

and put the faces where they're supposed to be on the hyperplanes, we end up with a polyhedral decomposition of the space. This also has the property that every vertex is in the closure of all the regions that it is supposed to border according to configuration C. Note that it is possible in the degenerate case for there to be some of the regions of our new decomposition that have smaller dimension than the entire space, and thus we might also have regions coinciding with each other. However, this will not be a problem, as when this occurs, we will still be able to split up the vertices into sets each corresponding to a region.

The important thing is that the polyhedral decomposition has the property that, for every region, its closure contains at least $n/2^Q$ points. In fact, we can go further and say that for any set T of t regions, the union of their closures contains at least $tn/2^Q$ points. Then by Hall's marriage Theorem, there exists a way of associating disjoint sets of $\lfloor n/2^Q \rfloor$ points to each region such that the points are inside the closure of that region. There are at most $2^Q - 1$ points left over, which we put in whichever region can accept them.

We have therefore created a partition of the points of V into 2^Q regions, such that each region contains between between $\lfloor n/2^Q \rfloor$ and $\lfloor n/2^Q \rfloor + 2^Q - 1$ of the points, and such that every hyperplane in $\mathbb{R}^Q$ must avoid at least one of the interiors of a region. Thus the lemma is proved.

□

The function $F^{(n)}$ that we construct will do two things: Given a vertex i, it will provide an "address" $A(i)$ that will make it easier to find. Secondly, it will provide a tree structure $B(i)$ which defines which addresses it has an edge to and which ones it doesn't. This takes the form of a tree with labels on all its nodes.

The address Apply Lemma 12 to split the space into 2^Q regions, each of which contains at most $\lfloor n/2^Q \rfloor + 2^Q - 1$ vertices. We'll number these regions $1, 2, ..., 2^Q$ and for each vertex x, we will then store the information about which region it is in as the first line of the address: $A_1(x)$. This takes Q bits per vertex.

Then repeat this process with every region, splitting each further into 2^Q subregions,

then splitting each subregions into 2^Q subsubregions, etc. Continue until there are only 4^Q vertices in any given subregion. This will end in a number of steps $s = \lceil \log_2 \left(\frac{n-2^Q+1}{4^Q-2^Q+1} \right) / Q \rceil = \lceil \log_2 \left(\frac{n}{4^Q} \right) / Q + \log_2 \left(\frac{1-2^Q/n+1/n}{1-2^{-Q}+2^{-2Q}} \right) / Q \rceil$ steps. Now since $Q \geq 1$, $1-2^{-Q}+2^{-2Q} \geq 3/4$ and $1-2^Q/n+1/n \leq 1$ so $\log_2 \left(\frac{1-2^Q/n+1/n}{1-2^{-Q}+2^{-2Q}} \right) / Q \leq \log_2(4/3) < 1$. So overall we get the number of steps s is at most $\lceil \log_2 \left(\frac{n}{4^Q} \right) / Q + 1 \rceil = \lceil \frac{\log_2(n)}{Q} - \frac{2Q}{Q} + 1 \rceil \leq \frac{\log_2(n)}{Q}$ We then split this final subregion into its constituent points. Since there are at most 4^Q vertices in this subregion, this final decomposition also only takes $2Q$ bits.

Thus, each vertex x has a unique address $A(x)$ which takes the form of a string of $s+2$ numbers: $(i_1, i_2, ..., i_{s+2})$ where each i_w is an integer between 1 and 2^Q. The total amount of information stored in each vertex for the address ends up being $Q(s+2) \leq \log_2(n)+2Q$.

The tree-structure Given a vertex y, we will construct the labelled tree $B(y)$ by induction. At step 0, we start with just the root node and leave it without a label. Throughout the construction, all the nodes in the tree can be matched onto certain partial addresses. The root node gets matched onto the empty address.

Suppose we are at a certain step of the algorithm and that there exists an unlabelled leaf node in the tree. Say it can be matched to the partial address $(i_1, i_2, ..., i_l)$. The first thing we do is give it 2^Q child nodes. We will match each of these child nodes to the addresses $(i_1, i_2, ..., i_l, t)$ for every value of t between 1 and 2^Q. Now because the graph is semi-algebraic, we know that there exists some half-space such that for every other vertex x, x is connected to y if and only if x is in that half-space. This half-space has a hyperplane as its boundary, which we'll call $\mathcal{H}$. Remember that when writing the address, we split the region $(i_1, i_2, ..., i_l)$ into 2^Q subregions using Lemma 12, so we know that $\mathcal{H}$ must avoid at least one of the interior's of a subregion. We will write a list of all the subregions whose interior it avoids on node $(i_1, i_2, ..., i_l)$. Say it avoids the interior of the tth subregion. Now this subregion's interior is either entirely contained within the half-space or it is entirely disjoint from it. In other words, either all the vertices in the interior of the subregion are adjacent to y or none of them are. In fact, because we assumed that $f(x, y) \neq 0$, for all x, y, $\mathcal{H}$ won't pass through any of the vertices, so this also extends to vertices on the boundary of the subregion. So we know that either all vertices in the region are adjacent to y or none of them are. If all the vertices are adjacent, we will write a "1" on the tth child node. Otherwise write a "0" on the tth child node. Leave all the other child nodes unlabelled for the time being.

Continue in this fashion until the only empty nodes in the tree correspond to sets of size less than 4^Q. This will eventually happen at step number s. For each of these empty nodes, write down the size of the corresponding set on the node, and then creates 4^Q child nodes, each with a '1' or a '0' to indicate whether it is or isn't adjacent to y. Thus, we will end up with a tree of depth at most $s+1$. This tree is comprised of some nodes with 2^Q child nodes; call these "splitting nodes" (except the final splitting nodes which have 4^Q children instead). The rest of the nodes just have a single number, "0" or "1" on them. We call these "leaf nodes".

The function G The function G is simple to construct. Given two vertices x and y, look at x's address. Say it is $(i_1, i_2, ..., i_{l+2})$. Now look at y's tree. Travel through this tree by starting at the root node, and at every step l, if we are at a splitting node, then go to the i_lth child node. Eventually we will reach a leaf node and at that point, we should be able to read "1" or "0". If there is a "1", that means there is an edge between x and y. If there is a "0", that means there is not.

Information used What is the maximum amount of information required to store this tree? The structure (whether a certain node has a child or not) is entirely determined by the numbers written on each node, so we only have to count up the total information stored in the numbers. We'll work backwards from the end.
Each leaf node has either a "0" or a "1" so we have 1 bit per leaf node.
The final splitting nodes at the end have at most 4^Q children, each requiring 1 bit, so that's 4^Q bits for the children. It also stores how many children it has which requires an additional $2Q$ bits. So $4^Q + 2Q$ bits suffice to store a splitting node at depth l with all its descendants.

Let $\alpha = 8^Q - 2 \cdot 4^Q + 2 \cdot 2^Q - 3Q - 1$. We will prove by induction that $\frac{\alpha(2^Q-1)^m-(Q+1)}{2^Q-2}$ bits suffices to to store a splitting node at depth $s-m$ together with all its descendants. When $m = 0$, it's easy to check our choice of α makes this hold.

Now suppose we have a splitting node **i** which is at depth $s-m$ for some $m \geq 1$. How much information suffices for it and all its descendants? Suppose it has a leaf nodes

adjacent. Each of these uses 1 bit for itself, and another Q bits to be put on the list of leaf nodes at $\mathbf{i}$, for a total of $a(Q+1)$ bits. The other $2^Q - a$ nodes are all splitting nodes, so by the induction hypothesis, each can be described using only $\frac{\alpha(2^Q-1)^{m-1}-(Q+1)}{2^Q-2}$ bits. Totalling everything up, we get: $\frac{\alpha(2^Q-1)^{m-1}-(Q+1)}{2^Q-2}(2^Q - a) + a(Q+1)$. Since $a \geq 1$, we get that this is less than: $\frac{\alpha(2^Q-1)^m}{2^Q-2} - \frac{(Q+1)(2^Q-1)}{2^Q-2} + (Q+1) = \frac{\alpha(2^Q-1)^m}{2^Q-2} - \frac{Q+1}{2^Q-2}$.

Therefore by induction, each splitting node at depth $s-m$ together with all its descendants can be described using only $\frac{\alpha(2^Q-1)^m}{2^Q-2} - \frac{Q+1}{2^Q-2}$ bits. Therefore the total number of bits that suffices to store the entire tree is $\frac{\alpha(2^Q-1)^l}{2^Q-2} - \frac{Q+1}{2^Q-2}$

Summing it all up:

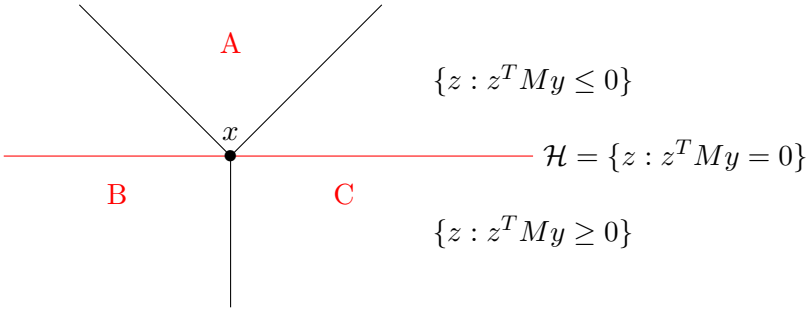
Summing the contribution from the address $A(y)$ and the contribution from the tree $B(y)$, we get that the total maximum number of bits that suffices to store $F^{(n)}(y)$ is:

$$\begin{aligned}
& Q(s+2) + \frac{\alpha(2^Q-1)^s}{2^Q-2} - \frac{Q+1}{2^Q-2} \\
\leq\ & \frac{\alpha(2^Q-1)^{\left(\frac{\log_2(n)}{Q}\right)}}{(2^Q-2)}(1+o(1)) \\
=\ & \frac{\alpha}{(2^Q-2)} 2^{\log_2(2^Q-1)\left(\frac{\log_2(n)}{Q}\right)}(1+o(1)) \\
=\ & n^{\log_2(2^Q-1)/Q}\left(\frac{\alpha}{(2^Q-2)} + o(1)\right) \\
=\ & n^{(1-\frac{1}{Q2^Q})(1+o(1))}\left(\frac{\alpha}{(2^Q-2)} + o(1)\right) \\
=\ & n^{(1-\frac{1}{Q2^Q})(1+o(1))}
\end{aligned}$$

When n is large, this is an improvement over the trivial upper bound of $\lceil (n-1)/2 \rceil + \lceil \log_2(n) \rceil$.

5.4.2 The case where $f(x, y) = 0$

The reason the previous method might not work in this case is that x will be on the hyperplane $\mathcal{H}$ corresponding to y, and it could happen that x is on the boundary of its region, and that $\mathcal{H}$ is tangent to it. Then the interior of the region containing x would be completely on one side of $\mathcal{H}$, so if we used that method, we might erroneously get information about the edge x, y.



Example of a possible problem: $\mathcal{H}$ is the hyperplane corresponding to y, A is a region labelled as negative, B and C are regions that will be subdivided, the vertex x is on $\mathcal{H}$ and thus adjacent to y but is counted as being in region A, and thus erroneously counted as non-adjacent to y.

The way we fix this is we will consider the boundaries of regions to be full regions themselves, each of which will get their own addresses. However, the key thing to note here is that every boundary region will have smaller dimension. More formally, start with the closures of the 2^Q regions of the original decomposition. If two regions intersect, then their intersection gets subtracted from both of the original regions, and is instead counted as a region of its own. Repeat this process until there are no more intersections. Since there were 2^Q parts originally, there are at most $2^{2^Q} - 1$ regions in the new decomposition. Also for every d between 1 and Q, there are at most $\binom{2^Q}{1+Q-d}$ regions of dimension d.

Address When we store the address $A(x)$ of a point x, we might need to write down some of these new boundary regions in the address if x happens to be in one of them. However, we claim that the address can still be written using only $2\lceil \log_2(n) \rceil + 2Q^2$ bits. Indeed, every time we have a region of dimension Q with n points in it, we subdivide it into 2^Q subregions of dimension Q that contain at most $n/2^Q$ points and for every d

between 1 and $Q-1$, at most $\binom{2^Q}{1+Q-d}$ subregions of dimension d, each of which contains at most n points. There is a single region of dimension 0: the center of the decomposition, which obviously contains at most 1 point.

When $n \leq 2^Q$, then we can easily decompose using only $\lceil \log_2(n) \rceil$ bits, which is well within the bound (by a factor of 2). For n is larger, there are 4 cases:

Case 1: subregions of dimension Q: For points in the subregions of dimension Q, we use the induction hypothesis to say that the last part of the addresses can be written in $2\lceil \log_2(n/2^Q) \rceil + 2Q^2$ bits. As there are 2^Q such subregions, we can indicate which one they are in using an additional Q bits. Finally, we use $\lceil \log_2(Q) \rceil$ bits at the start to indicate what d is. Therefore their full addresses can be written using $2\lceil \log_2(n/2^Q) \rceil + 2Q^2 + Q + \lceil \log_2(Q) \rceil = 2\lceil \log_2(n) \rceil - 2Q + 2Q^2 + Q + \lceil \log_2(Q) \rceil = 2\lceil \log_2(n) \rceil + 2Q^2 - (Q - \lceil \log_2(Q) \rceil)$ bits. Since $Q \geq \lceil \log_2(Q) \rceil$, this works.

Case 2: subregions of dimension d, where $1 \leq d \leq Q-1$: By the induction hypothesis, the last part of the addresses of points in the subregions of dimension d can be written using $2\lceil \log_2(n) \rceil + 2d^2$ bits. As there are at most $\binom{2^Q}{1+Q-d}$ such subregions, we can write identify which one it is using $Q(1+Q-d)$ bits. Finally, we use $\lceil \log_2(Q) \rceil$ bits to indicate what d is. Therefore we can write the addresses of these points using $2\lceil \log_2(n) \rceil + 2d^2 + Q(1+Q-d) + \lceil \log_2(Q) \rceil = 2\lceil \log_2(n) \rceil + 2Q^2 + (2d^2 - Qd - Q^2 + Q + \lceil \log_2(Q) \rceil)$ bits. The worst case scenario for d is either when $d = 1$ or when $d = Q-1$.

When $d = 1$, we have the number of bits is $2\lceil \log_2(n) \rceil + 2Q^2 + (2 - Q^2 + \lceil \log_2(Q) \rceil)$. But now for this case to even appear, we need, $Q \geq 2$ so $Q^2 \geq 2 + \lceil \log_2(Q) \rceil$) so this works.

When $d = Q-1$, we have the number of bits is $2\lceil \log_2(n) \rceil + 2Q^2 + (-2Q + 2 + \lceil \log_2(Q) \rceil)$. But as before, $Q \geq 2$ so $2Q - 2 \geq \lceil \log_2(Q) \rceil$) so this works.

Case 3: subregions of dimension 0 This only happens if x is directly at the center of the decomposition. Then we don't need any extra information to identify x. We only need $\lceil \log_2(Q) \rceil$ bits to indicate that $d = 0$. This is well below $2\lceil \log_2(n) \rceil + 2Q^2$ so this easily works.

Therefore, by induction we can write down the new address of every vertex x using $2\lceil \log_2(n) \rceil + 2Q^2$ bits regardless of what subregions x is in.

Tree Structure We also need to remake the tree structure $B(y)$ using similar methods. Instead of $2^Q - 1$ branches at every splitting node, we'll end up with $1 + \sum_{d=1}^{Q} \binom{2^Q}{1+Q-d}$ branches. However, all the new branches will have far less information on them, with the end result that the entire tree doesn't require that much more information to store. More precisely, we can store the tree for n vertices in dimension Q in $F(Q, n) = c_Q(2^Q - 1)^{\log_2(n)/Q} - c'_Q(2^{Q-1} - 1)^{\log_2(n)/(Q-1)}$ bits for some sufficiently large c'_Q and c_Q.

For the subregions of dimension Q, we do a similar thing to last time . We know that there is at least one subregion of dimension Q that avoids the hyperplane $\mathcal{H}$ corresponding to y, which means all the elements of this subregion are either all adjacent or all non-adjacent from y. We'll say that there are in fact $a \geq 1$ subregions of dimension Q that avoid $\mathcal{H}$. We can identify each one using Q bits, and then we need to add 1 more bit to say whether all elements are adjacent or non-adjacent to y. For all the others, we know that since each contains at most $n/2^Q$ points, by the induction hypothesis that can store all the information using $F(Q, n/2^Q)$ bits. In total, we can store this information using $a(Q + 1) + (2^Q - a)F(Q, n/2^Q)$ bits. Since $F(Q, n/2^Q) > Q + 1$, the worst case scenario is when $a = 1$.

For the subregions of smaller dimension, we know that there are less than 2^{2^Q} of them. Therefore we can identify each one using 2^Q bits. Each of these also needs its internal information storing. By the induction hypothesis, we can store each one using $F(d, n)$ bits where d is its dimension. Since $d \leq Q - 1$, we we know that this is less than $F(Q - 1, n)$. In total, we can store all the information about these subregions using $2^{2^Q}(2^Q + F(Q - 1, n))$ bits. Adding this all up, we get that we can store all the information about a region of dimension Q with n points in it using information:

$$
\begin{aligned}
& (Q+1) + (2^Q-1)F(Q, n/2^Q) + 2^{2^Q}(2^Q + F(Q-1,n)) \\
= \;& (Q+1) + c_Q(2^Q-1)(2^Q-1)^{\log_2(n/2^Q)/Q} \\
& -c'_Q(2^Q-1)(2^{Q-1}-1)^{\log_2(n/2^Q)/(Q-1)} + 2^{2^Q+Q} \\
& +c_{Q-1}2^{2^Q}(2^{Q-1}-1)^{\log_2(n)/(Q-1)} - c'_{Q-1}2^{2^Q}(2^{Q-2}-1)^{\log_2(n)/(Q-2)} \\
\leq \;& (Q+1+2^{2^Q+Q}) + c_Q(2^Q-1)^{\log_2(n)/Q} + \\
& \left[c_{Q-1}2^{2^Q} - c'_Q\frac{(2^Q-1)}{(2^{Q-1}-1)^{Q/(Q-1)}}\right](2^{Q-1}-1)^{\log_2(n)/(Q-1)} \\
\leq \;& c_Q(2^Q-1)^{\log_2(n)/Q} + \\
& \left[\left(Q+1+2^{2^Q+Q}+c_{Q-1}2^{2^Q}\right) - c'_Q\frac{(2^Q-1)}{(2^{Q-1}-1)^{Q/(Q-1)}}\right](2^{Q-1}-1)^{\log_2(n)/(Q-1)}
\end{aligned}
$$

If this ends up being less than or equal to $c_Q(2^Q-1)^{\log_2(n)/Q} - c'_Q(2^{Q-1}-1)^{\log_2(n)/Q-1}$, then the induction would be complete. This is equivalent to:

$$\left[\left(Q+1+2^{2^Q+Q}+c_{Q-1}2^{2^Q}\right) - c'_Q\frac{(2^Q-1)}{(2^{Q-1}-1)^{Q/(Q-1)}}\right] \leq -c'_Q$$

which is again equivalent to

$$c'_Q\left(\frac{(2^Q-1)}{(2^{Q-1}-1)^{Q/(Q-1)}} - 1\right) \geq Q+1+2^{2^Q+Q}+c_{Q-1}2^{2^Q}.$$

But now $\frac{(2^Q-1)}{(2^{Q-1}-1)^{Q/(Q-1)}} > 1$ because $\frac{(2^Q-1)^{Q-1}}{(2^{Q-1}-1)^Q} > 1$ because $(2^Q-1)^{Q-1} > (2^{Q-1}-1)^Q$. Therefore we can pick a $c'_Q = \frac{Q+1+2^{2^Q+Q}+c_{Q-1}2^{2^Q}}{\frac{(2^Q-1)}{(2^{Q-1}-1)^{Q/(Q-1)}}-1}$ and it will work.

Then we can pick c_Q large enough to make the initial conditions of the induction hold (i.e.: when $n < 2^Q$, we can describe everything using $F(Q,n)$ bits). If we do this too, then by induction, we'll get for all n, we can describe everything using $F(Q,n)$ bits.

Then by induction on Q, we'll have a series of numbers c_Q and c'_Q such that we can describe the entire tree using $c_Q(2^Q-1)^{\log_2(n)/Q} - c'_Q(2^{Q-1}-1)^{\log_2(n)/(Q-1)}$ bits.

This completes the proof. We have a method of storing the information about the edges which uses $O(n^{\log_2(2^Q-1)/Q})$ bits for every vertex.

Examples

For the category of unit disk graphs, we have $Q = 4$ which means this takes $O(n^{0.976723})$ bits per vertex.

For the category of disk graphs, we have $Q = 5$, which means this takes $O(n^{0.990839})$ bits per vertex.

These are very close to the trivial upper bound of $\lceil \frac{n-1}{2} \rceil + \log_2(n)$ but are still a small improvement over it when n is large.

Bibliography

[1] Noga Alon. Tools from higher algebra. In *Handbook of combinatorics*, pages 1749–1783. Elsevier, Amsterdam, 1995.

[2] Noga Alon, János Pach, Rom Pinchasi, Radoš Radoičić, and Micha Sharir. Crossing patterns of semi-algebraic sets. *Journal of Combinatorial Theory, Series A*, 111(2):310–326, 2005.

[3] Aistis Atminas, Andrew Collins, Vadim Lozin, and Victor Zamaraev. Implicit representations and factorial properties of graphs. *Discrete Mathematics*, 338(2):164–179, 2015.

[4] Pavle VM Blagojević, Boris Bukh, and Roman Karasev. Turán numbers for $K_{s,t}$-free graphs: Topological obstructions and algebraic constructions. *Israel Journal of Mathematics*, 197(1):199–214, 2013.

[5] George R Blakley and Prabir Roy. A Hölder type inequality for symmetric matrices with nonnegative entries. *Proceedings of the American Mathematical Society*, 16(6):1244–1245, 1965.

[6] Béla Bollobás and Tom Eccles. Partial shadows of set systems. *Combinatorics, Probability and Computing*, 24(05):825–828, 2015.

[7] Heinz Breu and David G Kirkpatrick. Unit disk graph recognition is NP-hard. *Computational Geometry*, 9(1-2):3–24, 1998.

[8] Boris Bukh and David Conlon. Rational exponents in extremal graph theory. *Journal of the European Mathematical Society*, 20(7):1747–1757, 2018.

[9] Paul Erdős. On extremal problems of graphs and generalized graphs. *Israel Journal of Mathematics*, 2(3):183–190, 1964.

[10] Paul Erdős. On the combinatorial problems which I would most like to see solved. *Combinatorica*, 1(1):25–42, 1981.

[11] Paul Erdős and Richard Rado. Intersection theorems for systems of sets. *Journal of the London Mathematical Society*, 1(1):85–90, 1960.

[12] Paul Erdős and Arthur H Stone. On the structure of linear graphs. *Bull. Amer. Math. Soc*, 52(1087-1091):1, 1946.

[13] Péter Frankl. All rationals occur as exponents. *Journal of Combinatorial Theory, Series A*, 42(2):200–206, 1986.

[14] Meinolf Geck. *An introduction to algebraic geometry and algebraic groups*. Oxford University Press, 2013.

[15] Ross J Kang and Tobias Müller. Sphere and dot product representations of graphs. *Discrete & Computational Geometry*, 47(3):548–568, 2012.

[16] Sampath Kannan, Moni Naor, and Steven Rudich. Implicit representation of graphs. *SIAM Journal on Discrete Mathematics*, 5(4):596–603, 1992.

[17] Gyula Katona. A theorem of finite sets. In *Theory of graphs (Proc. Colloq., Tihany, 1966)*, pages 187–207. Academic Press, New York, 1968.

[18] Peter Keevash. Hypergraph Turán problems. *Surveys in combinatorics*, 392:83–140, 2011.

[19] Joseph B Kruskal. The number of simplices in a complex. *Mathematical optimization techniques*, page 251, 1963.

[20] Serge Lang and André Weil. Number of points of varieties in finite fields. *American Journal of Mathematics*, 76(4):819–827, 1954.

[21] Jie Ma, Xiaofan Yuan, and Mingwei Zhang. Some extremal results on complete degenerate hypergraphs. *arXiv preprint arXiv:1612.01363*, 2016.

[22] Colin McDiarmid and Tobias Müller. Integer realizations of disk and segment graphs. *Journal of Combinatorial Theory, Series B*, 103(1):114–143, 2013.

[23] Alexander Sidorenko. Inequalities for functionals generated by bipartite graphs. *Diskretnaya Matematika*, 3(3):50–65, 1991.

[24] Alexander Sidorenko. A correlation inequality for bipartite graphs. *Graphs and Combinatorics*, 9(2-4):201–204, 1993.

[25] Alexander Sidorenko. What we know and what we do not know about Turán numbers. *Graphs and Combinatorics*, 11(2):179–199, 1995.

[26] Miklós Simonovits. Extremal graph problems, degenerate extremal problems, and supersaturated graphs. *Progress in graph theory (editied by J. Adrian Bondy and U. S. R. Murty.)(Waterloo, Ont., 1982), Academic Press, Toronto, ON*, pages 419–437, 1984.

[27] Jeremy P Spinrad. *Efficient graph representations*. American Mathematical Society, 2003.

[28] Balazs Szegedy. An information theoretic approach to Sidorenko's conjecture. *arXiv preprint arXiv:1406.6738*, 2014.

[29] Paul Turán. On an extremal problem in graph theory. *Mat. Fiz. Lapok*, 48(436-452):137, 1941.

[30] Hugh E. Warren. Lower bounds for approximation by nonlinear manifolds. *Transactions of the American Mathematical Society*, 133(1):167–178, 1968.

[31] Andrew C Yao and F Frances Yao. A general approach to d-dimensional geometric queries. In *Proceedings of the seventeenth annual ACM symposium on Theory of computing*, pages 163–168. ACM, 1985.